CH1

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 November 2000 (30.11.2000)

PCT

(10) International Publication Number
WO 00/72182 A2

(51) International Patent Classification[7]: G06F 17/00

(21) International Application Number: PCT/US00/14159

(22) International Filing Date: 23 May 2000 (23.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/318,699 25 May 1999 (25.05.1999) US

(71) Applicant *(for all designated States except US)*: **DIGITAL GENE TECHNOLOGIES, INC.** [US/US]; 11149 North Torrey Pines Road, Suite 110, La Jolla, CA 90237 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **GRACE, Dennis, R.** [US/US]; 3137 Fenelon Street, San Diego, CA 92106 (US). **DURHAM, Jayson, T.** [US/US]; 10359 Mountain View Lane, Lakeside, CA 92040 (US).

(74) Agent: **LESAVICH, Stephen**; McDonnell Boehnen Hulbert & Berghoff, Suite 3200, 300 South Wacker Drive, Chicago, IL 60606 (US).

(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 00/72182 A2

(54) Title: METHODS AND SYSTEM FOR AMPLITUDE NORMALIZATION AND SELECTION OF DATA PEAKS

(57) **Abstract:** Methods and system for amplitude normalization and selection of data peaks from experimental data including polynucleotide data such as DNA, cDNA or mRNA from biotechnology experiments. The methods and systems include removing spectral overlap, spatially detrending and normalizing a multi-component data signal into experimental data from a desired experiment (e.g., biotechnology data). Standard data sizes are determined and data clutter is rejected for filtered experimental data. Data sizes are calibrated and error removed from experimental data. Data stutter is removed and the number of data values is reduced. The methods and system help automate the processing of experimental data to eliminate or reduce errors and leave processed experimental data in a format suitable for visual display, comparative analysis and other analysis. The methods and systems may help reduce or eliminate inconsistencies in processing experimental data that typically lead to unreliable or erroneous results. The methods and system of the present invention may be used to refine processing of biotechnology data with new techniques that can be used for bioinformatics and for other types of experimental data that are visual displayed (e.g., telecommunications data, electrical data for electrical devices, optical data, physical data, or other data).

METHODS AND SYSTEM FOR AMPLITUDE NORMALIZATION AND SELECTION OF DATA PEAKS

FIELD OF THE INVENTION

This invention relates to selecting and processing data points from experimental data. More specifically, it relates to methods for amplitude normalization and selection of data peaks from experimental data, such as biotechnology data.

BACKGROUND OF THE INVENTION

Biotechnology data is collected and analyzed for many diverse purposes. As is known in the art, biotechnology data typically includes data obtained from biological systems, biological processes, biochemical processes, biophysical processes, or chemical processes. For example, sequences of deoxyribonucleic Acid ("DNA") from many different types of living organisms are often determined and mapped. DNA is double-stranded heteropolymer including a continuous string of four nucleotide base elements. The four nucleotide base elements include deoxyadenosine, deoxycytidine, deoxyguanosine, and deoxythymidine. The four nucleotide bases are usually abbreviated as "A," "C," "G" and "T" respectively. DNA is used to make ribonucleic acid ("RNA"), which in turn is used to make proteins. "Genes" are regions of DNA that are transcribed into RNA, which encodes a translated protein.

One fundamental goal of biochemical research is to map and characterize all of the protein molecules from genes in a living organism. The existence and concentration of protein molecules typically help determine if a gene is "expressed" or "repressed" in a given situation. Protein characterization includes, identification, sequence determination, expression, characteristics, concentrations and biochemical

activity. Responses of proteins to natural and artificial compounds are used to develop new treatments for diseases, improve existing drugs, develop new drugs and for other medical applications.

Biotechnology data is inherently complex. For example, DNA sequences include large numbers of A's, C's, G's and T's, that need to be stored and retrieved in a manner that is appropriate for analysis. "Bioinformatics" techniques are used to address problems related to biotechnology information (e.g., DNA, RNA and protein sequences). As is known in the art, bioinformatics is the systematic development and application of information technologies and data mining techniques for processing, analyzing and displaying data obtained by experiments, modeling, database searching, and instrumentation to make observations about biological processes.

There are a number of problems associated with using biotechnology data. One problem is the collection of biotechnology data. Biotechnology data is commonly presented as graphical plots of two or more variables. A "peak," i.e., a local maximum in a plot of two or more variables, is often a feature of interest in biotechnology data. When biotechnology data is collected, the collection process often introduces reference signals and/or error signals that are based on equipment used to collect the data and the calibration of the equipment. For example, DNA sequences may be determined by processing samples using gel-electrophoresis. A label (e.g., a dye) is incorporated into the samples for detection by laser-induced fluorescence. Electrophoresis resolves molecules from the samples into distinct bands of measurable lengths on a gel plate. The gel-electrophoresis process typically provides biotechnology data that includes not only desired sequence information, but also reference and/or error signal information from the gel-electrophoresis process.

The reference signals have to be determined and the error signals removed to provide accurate biotechnology data.

Another problem is that collected biotechnology data has to be processed after collection so it is stored in a useable format. Even after any reference signals have been determined and any error signals removed, the remaining biotechnology data typically is still not in an appropriate format for storage. For example, the biotechnology data may have to be normalized to allow comparisons to other sets of collected data, scaled to appropriate values for display, filtered to remove unwanted data, or otherwise processed so all collected data is stored and retrieved in a uniform manner.

Yet another problem is that biotechnology data that is processed and stored may not be appropriate for visual display. As is known in the art, one of the most commonly used methodologies in biotechnology is "comparison." Many biological objects are associated with families that share the same structural or functional features. For example, many proteins with a similar sequence may have common functionality.

Visual display of biotechnology data is typically recognized as typically being "necessary" for biotechnology research. Visual display tools allow creation of complex views of large amounts of inter-related data, typically displaying information using computer generated graphics. For example, a discovered DNA sequence comprising a string of one thousand A's, C's, G's and T's is not easily compared to a known protein sequence with a hundreds of amino acids as a textual string. A new protein sequence discovered in an experiment may be more easily visually compared to known protein sequences using multi-colored visual displays. In addition, a protein sequence may be displayed visually in three-dimensions to aid analysis. Visual

display of biotechnology data typically requires additional processing of experimental data to allow the data to be visually displayed with an appropriate scale on a Graphical User Interface ("GUI") on a computer display.

Yet another problem is that the amount of biotechnology data collected from an experiment is typically large and the data is often processed by hand, typically by lab technicians. This often slows processing of the data and can introduce inconsistencies into the data (e.g., as a result of varying levels of skill and experience of the lab technicians). The inconsistencies in the data processing may lead to unreliable or erroneous results.

Yet another problem is that selecting appropriate subset of biotechnology data to display, from a large set of biotechnology data involves "combinatorics." As is known in the art, combinatorics relates to the arrangement of, operation on, and selection of, discrete elements belonging to finite sets of data points. A "naive" or "brutal force" attempt to process combinatoric biotechnology data is typically beyond the capabilities of the current generation of computers.

Yet another problem is that biotechnology data may include legitimate or outlying data points that may skew experimental results. Applying many of bioinformatic techniques known in the art to process biotechnology data may produce wild data points or discontinuities that affect visual display and analysis of the data.

Thus, it is desirable to process experimental biotechnology data and other types of data collected with a variety of laboratory equipment to produce data that can be reliably stored, retrieved, and visually analyzed. It is also desirable to automate processing of biotechnology data to provide timely and reproducible interpretations of biotechnology data without introducing inconsistencies into the data.

SUMMARY OF THE INVENTION

In accordance with preferred embodiments of the present invention, some of the problems associated with processing experimental data are overcome. Methods and system for amplitude normalization and selection of data peaks is provided. One aspect of the present invention includes a method for removing spectral overlap, spatially detrending, and normalizing a multi-component data signal into filtered experimental data from a desired experiment (e.g., biotechnology data). The multi-component signal could include, but is not limited to, laser-induced fluorescence of biotechonology data.

Another aspect of the present invention includes a method for detection of data standards and data clutter rejection from filtered experimental data to create processed experimental data. Another aspect of the present invention includes a method for data size calibration and error removal from processed experimental data. Another aspect of the present invention includes data stutter removal and reduction of the number of processed experimental data values. Another aspect of the present invention includes a general method for processing general multi-component data signals including biotechnology data. Another aspect of the present invention includes a system for processing experimental data into a format that is suitable for display on a display device for comparative and other analysis.

The methods and system described herein help automate the processing of experimental data to eliminate or reduce errors and leave processed experimental data in a format suitable for visual display, comparative analysis or other analysis. The methods and systems may help reduce or eliminate inconsistencies in processing experimental data that typically lead to unreliable or erroneous results. The methods and system of the present invention may be used to refine processing of

biotechnology data with new techniques that can be used for bioinformatics and for other types of experimental data that are visual displayed (e.g., telecommunications data, electrical data for electrical devices, optical data, physical data).

The foregoing and other features and advantages of preferred embodiments of the present invention will be more readily apparent from the following detailed description. The detailed description proceeds with references to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of the present invention are described with reference to the following drawings, wherein:

FIG. 1 is a block diagram illustrating an exemplary experimental data processing system;

FIG. 2 is a flow diagram illustrating a method for data normalization for a multi-component data signal;

FIG. 3A is a block diagram illustrating an exemplary unfiltered signal intensity trace for a multi-component data signal;

FIG. 3B is a block diagram illustrating the unfiltered multi-component data signal of FIG. 3A as an unfiltered multi-component data signal displayed with a larger scale;

FIG. 3C is a block diagram illustrating a filtered version of the multi-component data signal of FIG. 3A;

FIG. 3D is a block diagram illustrating a filtered and normalized multi-component data signal using the method from FIG. 2;

FIG. 4 is a flow diagram illustrating a method of clutter rejection;

FIG. 5 is a block diagram illustrating a filtered and normalized multi-component data signal using the method from FIG. 2;

FIG. 6 is a block diagram illustrating a filtered standard for a sequence of scans for a set of lanes in an electrophoresis-gel that were loaded with standard polynucleotide fragments at the same time;

FIG. 7 is a block diagram illustrating data peaks with size standard detection with clutter rejection using the method of FIG. 4;

FIG. 8 is a block diagram illustrating a method for data size calibration;

FIGS. 9A and 9B are block diagrams illustrating data size calibration using the method from FIG. 8;

FIG. 10 is a flow diagram illustrating a method for envelope detection;

FIGS. 11A and 11B are block diagrams illustrating envelope detection using the method of FIG. 10;

FIGS. 12A and 12B is a flow diagram illustrating a method for processing multi-component experimental data;

FIGS. 13A and 13B are block diagrams illustrating the method of FIGS. 12A and 12B; and

FIG. 14 is a block diagram illustrating an exemplary multi-component signal data processing system.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

In one exemplary preferred embodiment of the present invention, biotechnology data for simultaneous sequence-specific identification of expressed genes is processed with the methods and system described herewith. However, the present invention is not limited to processing biotechnology data, and methods and system described herein can be used to process other data (e.g., telecommunications data, electrical data, optical data, physical data, other data. etc.).

Gene Mapping

As was discussed above, deoxyribonucleic acid ("DNA") is a double-stranded heteropolymer that can be thought of symbolically as a continuous string of four nucleotide base elements, deoxyadenosine, deoxycytidine, deoxyguanosine, and deoxythymidine. The four bases are usually abbreviated as "A," "C," "G" and "T" respectively, and base elements on one strand of DNA interact with a counterpart on the other strand. For example, an "A" can only interact with a "T," and a "G" can only interact with a "C." This relationship is called "base pairing." "Genes" are regions of DNA, and "proteins" are the products of genes. Proteins are built from a fundamental set of amino acids, and DNA carries amino-acid coding information. When DNA is replicated or copied, a new DNA strand is synthesized using each of the original strands as templates.

DNA itself does not act as a template for protein decoding or synthesizing. A complementary copy of one of the two strands of DNA is synthesized out of ribose nucleotides to generate a ribonucleic acid ("RNA") copy of a gene with a method called "transcription." The RNA copy of a gene is then decoded by protein synthesis with a method called "translation." Since the RNA carries protein codes, it is called messenger RNA ("mRNA"). The transcription of mRNA is very precise and always

starts at one precise nucleotide and ends exactly at another. Complementary DNA ("cDNA") is an exact, double-stranded DNA copy of mRNA. One of the cDNA strands is complementary to the mRNA, and other is identical.

There are many techniques known in the biotechnology arts to identify RNA species including, but not limited to those described in "Differential display of eukaryotic messenger RNA by means of polymerase chain reaction," by P. Liang and A. B. Pardee, Science, Vol. 257, pages 967-971, 1992; "Arbitrarily primed PCR fingerprinting of RNA," by J. Welsh, K. Chada, S. S. Dalal, R. Cheng, D. Ralph and M. McCelland, Nucleic Acids Research, Vol. 20, pages 4965-4970, 1992; "A simple and very efficient method for generating cDNA libraries," Gene, Vol. 25, pages 263-269, 1983; "Tissue-specific expression of mouse α-amylase genes," by K. Schibler, M. Tosi, A.C. Pittet, L. Fabiani and P.K. Wellauer, Journal of Molecular Biology, Vol. 142, pages 93-116, 1990; "Discovering the secrets of DNA," by P. Friedland and L. H. Kedes, Communications of the Association for Computing Machinery ("CACM"), Vol. 28, No. 11, pages 1164-1186, November 1985; and others.

RNA isolated from a target organism (e.g., a cell to which a new drug has been applied) is analyzed using a method of simultaneous sequence-specific identification of mRNAs. In one preferred embodiment of the present invention, simultaneous sequence-specific identification of mRNAs is provided with a TOtal Gene expression Analysis method ("TOGA"), described in U.S. Patent No. 5,459,037 and U.S. Patent No. 5,807,680, incorporated herein by reference. However, other methods can also be used to provide sequence-specific identification of mRNAs, and the present invention is not limited to TOGA sequence-specific identification of mRNAs.

In one preferred embodiment of the present invention, preferably, prior to the application of the TOGA method or other methods, the isolated RNA is enriched to form a starting polyA-containing mRNA population by methods known in the art. In such a preferred embodiment, the TOGA method further comprises an additional Polymerase Chain Reaction ("PCR") step performed using one of four 5' PCR primers and cDNA templates prepared from a population of antisense complementary RNA ("cRNA"). A final PCR step using one of a possible 256 5' PCR primers and a universal 3' PCR primer produces as PCR products, cDNA fragments that corresponded to a 3'-region of the starting mRNA population.

A label (e.g., a dye) is incorporated in the PCR products to permit detection of the PCR products by laser-induced fluorescence. Gel-electrophoresis or equivalent techniques are used to resolve molecules from the PCR products into distinct bands of measurable lengths (See, e.g., FIG. 6). The produced PCR products can be identified by a) an initial 5' sequence comprising a nucleotide base sequence of a remainder of a recognition site or a restriction endonuclease that was used to cut and isolate a 3' region of cDNA reverse transcripts made from a mRNA population, plus the nucleotide base sequence of preferably four parsing bases immediately 3' to the remainder of the restriction enconuclease recognition site, or more preferably the sequence of the entire fragment; and b) the length of the fragment.

Processing PCR product data, including a nucleotide base sequence is a very complex task. Whether the TOGA method is used or not, the nucleotide sequences near the end of mRNA molecules give each mRNA an almost unique identity. In addition, data concerning a position and an amplitude of laser-induced fluorescence signals for PCR products are digitized and used to determine the presence and relative concentration of corresponding starting mRNA species. For example, PCR product

data is digitized by creating a data file with digital information. The data file may include digital values, for example, of optical brightness of electrophoresis patterns or other data used to identify the mRNA (e.g., data from a micro-array on a chip used to isolate the mRNA). To aid in the detection and analysis of mRNA sequences, a data file including experimental data is processed. In one exemplary preferred embodiment of the present invention, an experimental data processing system is used to process experimental data.

In one preferred embodiment of the present invention, the experimental data includes polynucleotide data for DNA, cDNA, cRNA, mRNA, or other-polynucleotides. The polynucleotide data can include, but is not limited to a length of a nucleotide fragment, a base composition of a nucleotide fragment, a base sequence of a nucleotide fragment, an intensity of a dye label signal used to tag a nucleotide fragment, or other nucleotide data. However, the present invention is not limited to polynucleotide data and other experimental data can also be used.

Exemplary experimental data processing system

FIG. 1 is a block diagram illustrating an exemplary experimental data processing system 10 for one exemplary preferred embodiment of the present invention. The experimental data processing system 10 includes a computer 12 with a computer display 14. The computer display 14 presents a windowed graphical user interface ("GUI") 16 to a user. A database 18 includes biotechnology experimental information or other experimental information. The database 18 may be integral to a memory system on the computer 12 or in secondary storage such as a hard disk, floppy disk, optical disk, or other non-volatile mass storage devices.

An operating environment for the data processing system 10 for preferred embodiments of the present invention include a processing system with one or more

high speed Central Processing Unit(s) ("CPU") and a memory. The CPU may be electrical or biological. In accordance with the practices of persons skilled in the art of computer programming, the present invention is described below with reference to acts and symbolic representations of operations or instructions that are performed by the processing system, unless indicated otherwise. Such acts and operations or instructions are referred to as being "computer-executed" or "CPU executed."

It will be appreciated that acts and symbolically represented operations or instructions include the manipulation of electrical signals or biological signals by the CPU. An electrical system or biological system represents data bits which cause a resulting transformation or reduction of the electrical signals or biological signals, and the maintenance of data bits at memory locations in a memory system to thereby reconfigure or otherwise alter the CPU's operation, as well as other processing of signals. The memory locations where data bits are maintained are physical locations that have particular electrical, magnetic, optical, or organic properties corresponding to the data bits.

The data bits may also be maintained on a computer readable medium including magnetic disks, optical disks, organic memory, and any other volatile (e.g., Random Access Memory ("RAM")) or non-volatile (e.g., Read-Only Memory ("ROM")) mass storage system readable by the CPU. The computer readable medium includes cooperating or interconnected computer readable medium, which exist exclusively on the processing system or be distributed among multiple interconnected processing systems that may be local or remote to the processing system.

Analyzing biotechnology data

In one exemplary preferred embodiment of the present invention, a label is incorporated in target biotechnology products (e.g., polynucleotide PCR products) for detection by laser-induced fluorescence and electrophoresis is used to obtain biotechnology data. However, other techniques may also be used to collect experimental biotechnology data (e.g., micro-arrays).

A complex, multi-component information signal based on indicated fluorescence intensities of the biotechnology products is included in a resulting experimental data file as digital data. The multi-component information signal includes raw multi-component label fluorescence intensities. Label responses are relatively broadband spectrally and typically include spectral overlap. Energy measured as a second fluorescence response typically includes energy in the tail of a first fluorescence response, which might also be present, and vice-versa.

This spectral overlap needs to be removed because the relative quantities of commingled energy may be of a same order of magnitude as relative fluorescence responses of the data representing target data (e.g., polynucleotide data). For example, a small fluorescence response for a given polynucleotide data fragment in a biotechnology product may be "overwhelmed" if it occurs in a spectral overlap region between two fluorescence responses. In an exemplary preferred embodiment of the present invention, spectral overlap is removed and a normalized baseline is created with a combination of filtering techniques.

Removing spectral overlap and normalizing data

FIG. 2 is a flow diagram illustrating a Method 20 for data normalization of a multi-component data signal. At Step 22, a multi-component data signal is read. The multi-component data signal includes multiple individual data signal components of varying spectral characteristics with varying amplitudes. The multiple individual data

signal components overlap within portions of the multi-component data signal. At Step 24, a spectral filter is applied to the multi-component data signal to create multiple non-overlapping individual data signal components. At Step 26, a spatial filter is applied to multiple signal artifacts in the multi-component data signal that introduce ambiguity to base values in the multiple non-overlapping individual data signal components to spatially detrend and normalize the multiple non-overlapping individual data signal components to a uniform base value.

In one preferred embodiment of the present invention, the spectral characteristics of the multi-component data signal comprise physical attributes and conditions including but not limited to, an absorption spectrum of a dye label, an emission spectrum of a dye label, an emission wavelength power and pulse duration of an exciting laser, or other spectral characteristics. The spectral filtering at Step 24 of Method 20 includes "demultiplexing" or separating individual components of raw fluorescence intensities that are combined by overlap of spectral characteristics of different dyes used to tag polynucleotide data (e.g., DNA, mRNA or cDNA). A "dye taggant" is a substance added to data of interest that fluoresces under laser excitement. In one exemplary preferred embodiment of the present invention, polynucleotide data or other data is tagged with a dye taggant. However, Method 20 is not limited to processing fluorescence intensities from polynucleotide data and can be used to process other types of data that generate a multi-component data signal.

In one exemplary preferred embodiment of the present invention, spectral filtering makes use of a set of coefficients that represent a relative degree to which energy in fluorescence responses of various dye taggants overlap. Denoting this set of coefficients by $\{m(p,q)\}$, $m(p,q)$ is a measurement of an amount of energy measured at a wavelength that corresponds to a center of a fluorescence response of a **p-th** dye

taggant, which is actually due to fluorescence response of a **q-th** dye taggant at that wavelength. The total unfiltered fluorescence response measured at any such central wavelength is then taken to be a weighted sum of the actual dye-specific fluorescence response. An unfiltered. measured fluorescence intensity at the central wavelength of the **p-th** dye taggant is denoted as **A'(p)** and an actual dye-specific fluorescence intensity is denoted as **A(q)**. In terms of these conventions. Equation 1 illustrates a relationship between measured and actual fluorescence intensities.

$$A'(p) = \Sigma_q m(p,q)\, A(q) \qquad (1)$$

The spectral filter comprises extracting the actual fluorescence intensity **A(q)**, by inverting a linear system of equations in Equation 1 using a singular value decomposition of a coefficient matrix **m(p,q)**. The spectral overlap coefficients **m(p,q)** and unfiltered fluorescence intensity **A'(p)** are typically obtained from measurements as part of the calibration of instrumentation used to produce and record the fluorescence intensities. However, these values can also be obtained from other sources. This extraction is an exemplary spectral filter used at Step 24 of Method 20. However, other spectral filters could also be used and the present invention is not limited to the spectral filters illustrated by the inversion of Equation 1.

The spectral filter is followed by a spatial filter at Step 26 of Method 20. In one exemplary preferred embodiment of the present invention. the spatial filter is a nonlinear morphological gray-scale "rolling ball" transformation, which spatially detrends and normalizes the intensities to a set of uniform base line values. However, other types of spatial filters could also be used and the present invention is not limited to the spatial filters described herein.

In one exemplary preferred embodiment of the present invention, the nonlinear morphological gray-scale rolling ball transformation that spatially "detrends" and "normalizes" the fluorescence intensity traces to a set of uniform base line values has two stages. The first stage creates a version of a trace that excludes local variations whose spatial extent is below a certain scale. This scale is chosen to be slightly greater than a measured extent along a trace of typical standard data peaks, so a resulting trace very closely resembles an original trace with peaked regions on a spatial scale of standard peaks and smaller peaks smoothed away. In preferred embodiments of the present invention, data peaks include entities having at least two dimensions characterized by a maximum amplitude and a width. The data peaks may also be described by a width at a half-maximum amplitude or a position of a maximum amplitude.

This inherently nonlinear process is followed in a second stage by forming a difference between an original and a smoothed version of the trace, leaving a uniformly base-lined residual including peaked regions on a spatial scale of standard peaks and smaller. The term "rolling ball" refers to how the smoothed version of a trace is formed in a first stage of this filtering. In effect, a "ball" of a radius set by a exclusion scale of interest is first "rolled" along an under side of a trace, while maintaining at least one point of contact with the trace. A new trace is formed by taking, at each sample index (e.g., a scan line), a highest point of the ball when its center is on a sample index. This is followed by a pass of the same ball along the top side of this new trace, with a final new trace formed by taking, at each sample index, the lowest point of the ball when its center is on the sample index.

If $\mathbf{f(n)}$ is a fluorescence intensity of a trace measured at sample index $\mathbf{n}$, $\mathbf{f_{min}}$ is set equal to a minimum fluorescence intensity across an entire trace. A spatial scale

of standard peak features is taken to be slightly less than N-sample indices (e.g., N-scan lines). The trace is first "eroded" by forming a new trace $\mathbf{f_-(n)}$ as illustrated in Equation 2.

$$\mathbf{f_-(n) \equiv \min \{ f(n+m) - f_{min} : -N/2 \le m \le N/2 \}} \quad (2)$$

The eroded trace $\mathbf{f_-(n)}$ from Equation 2 is "dilated" as illustrated in Equation 3.

$$\mathbf{f_+(n) \equiv \max \{ f_-(n+m) + f_{min} : -N/2 \le m \le N/2 \}} \quad (3)$$

A fluorescence intensity of the rolling ball filtered version of an original trace at sample index $\mathbf{n}$ is $\mathbf{f_0(n)}$ as is illustrated in Equation 4.

$$\mathbf{f_0(n) \equiv f_-(n) - f_+(n)} \quad (4)$$

It is a sequence of finding minima and maxima (e.g., Equations 2 and 3) that accounts for the nonlinearility of the filter. Data values are normalized to a set of uniform base values.

The present invention with Method 20 is not limited to processing and normalizing biotechnology data multi-component signal or processing data with Equations 1-4 and can be used for other data from a multi-component signal (e.g., telecommunications signals, electrical signals data for electrical devices, optical signals, physical signals, or other data signals).

In one exemplary preferred embodiment of the present invention, "control" or "standard" polynucleotide data fragments (i.e., known polynucleotide data fragments) are tagged with a dye, which under laser illumination responds with a "red" fluorescence, while "target" polynucleotide data fragments (i.e., polynucleotide data to be identified) are tagged with a dye which has a "blue" response. However, the dyes used for the control and target could also be interchanged. Both the red and blue dye responses are relatively broadband spectrally, to the extent that energy

measured as red fluorescence response includes energy in a tail of any blue fluorescence response which might also be present and vice-versa. This spectral overlap is taken into account because the relative quantities of commingled energy are of the order of the relative fluorescence intensities of the target polynucleotide data and standard polynucleotide data fragments.

FIG. 3A is a block diagram 28 of an unfiltered multi-component data signal 30. FIGS. 3A-3D are used to illustrate use of Method 20 of FIG. 2. In one exemplary preferred embodiment of the present invention, the multi-component data signal 30 is a measurement of signal intensity of fluorescence on a vertical axis 32 at a fixed point in an electrophoresis-gel at successive points in time. The signal intensity of fluorescence is directly proportional to a parameter on a horizontal axis 34 representing a sample index (e.g., a scan line). However, other multi-component signal data could also be used and the present invention is not limited to polynucleotide fluorescence intensity data. A magnitude of the fluorescence intensity at a given scan line has been demonstrated to represent an amount of tagged polynucleotide fragments at a fixed point in time of a scan (e.g., tagged with red or blue dyes). The scale of standard polynucleotide fragment fluorescence intensity is illustrated by the narrow peak 36, of about two-hundred fluorescence units, which is illustrated in the region near sample index 2500 (e.g., 2500 scan lines) on the horizontal axis 34. In one preferred embodiment of the present invention, FIG. 3A illustrates a multi-component data signal 30 for a standard set of polynucleotide fragments.

FIG. 3B is a block diagram 38 illustrating the unfiltered multi-component data signal 30 for a standard set of polynucleotides fragments of FIG. 3A as an unfiltered multi-component data signal 40 displayed with a larger scale. FIG. 3C is a block

diagram 42 illustrating a filtered version of a multi-component data signal 44 for a target set of polynucleotides. The filtered version of the multi-component data signal 44 for the target set of polynucleotides (FIG. 3C) is at least an order of magnitude greater than that of the unfiltered multi-component data signal 40 for a standard set of polynucleotides (FIG. 3B).

A degree of spectral overlap is illustrated by the presence, in the unfiltered multi-component data signal 40 for a standard set of polynucleotides of FIG. 3B, of such artifacts as the broad peaks 46 in the region of sample index 2500 (e.g., 2500 scan lines) on the horizontal axis 32. The broad peaks 46 of FIG. 3B, when compared with the narrower peaks 48 of FIG. 3C, are due to spectral overlap of blue fluorescence intensities from blue-tagged target polynucleotide fragments since there are no red-tagged standard polynucleotide fragments that could produce such levels of flourescence intensities. An ambiguous baseline in this region (i.e., 2500 scan lines) illustrates "spectral bleed through" of blue-tagged target polynucleotide fragments that dramatically dwarf red-tagged standard polynucleotide fragments of interest.

FIG. 3D is a block diagram 52 illustrating application of Method 20 of FIG. 2 to the unfiltered multi-component data signal 30 for the standard set of polynucleotide fragments of FIG. 3A. FIGS. 3A and 3D use the same signal intensity scale to allow direct comparison. Note the clean data peaks 54, 56, 58, 60, 62, 64, 66, 68, 70 and 72 in FIG. 3D normalized to a uniform base value by applying the spectral and spatial filters of Method 20 to the unfiltered multi-component data signal 30 for the standard set polynucleotide fragments of FIG. 3A. Method 20 of FIG. 2 is also applied to the multi-component data signal for the target set of polynucleotides of FIG. 3B to produce set of clean peaks similar to those in FIG. 3D (this is not illustrated in FIG. 3).

Data standards size detection, error removal and clutter rejection

The multi-component data signals filtered and normalized to a baseline value with Method 20 of FIG. 2 may still contain false or erroneous data peaks due to false peak clutter. Such erroneous or false data peaks, if not removed, may skew experimental results. In one exemplary preferred embodiment of the present invention size standards detection with removal of false peak clutter rejection is used to identify a set of valid biotechnology fragment data from a filtered set of biotechnology fragment data (e.g., polynucleotide data). However, size standards detection with removal of false peak clutter can also be used on data other than biotechnology fragment data.

FIG. 4 is a flow diagram illustrating a Method 74 of clutter rejection. At Step 76, a first set of data points is selected from a filtered set of data points (e.g., filtered using Method 20, FIG. 2) using initial threshold criterion. At Step 78, multiple overlapping subsets of data points are selected from the first set of data points. At Step 80, multiple linear mappings are applied to the multiple overlapping subsets of data points. At Step 82, multiple error values are determined from the application of the multiple linear mappings to the multiple overlapping sub-set of data points. At Step 84, a first final subset of overlapping data points with a smallest error value is selected from the first set data points. Data points in the first final subset of overlapping data points include data points that fall within a standardized range where false data points have been removed.

In one exemplary preferred embodiment of the present invention, peaks in candidate biotechnology fragment data are located at Step 76 (FIG. 4) in filtered biotechnology fluorescence intensity data (e.g., with Method 20) using thresholds on simple ratios of differences between "microscale" and "mesoscale" average

fluorescence intensity levels relative to mesoscale variances. However, other thresholds could also be used.

There are typically a very large number of sets of filtered data points that can be selected for use with Method 74. Thus, selecting an appropriate filtered set of data points is a "combinatorics" problem. As was discussed above, combinatorics relates to the arrangement of, operation on, and selection of discrete elements belonging to finite sets of data points. However, Method 74 reduces the combinatorics of data selection to a "best" possible solution using multiple linear mappings, and allows a best set of data points (e.g. for a data peak mapping) to be created from a very large set of filtered data points. Method 74 provides an accurate selection of data points on data sub-scale, instead of a electrophoresis-gel scale, thus reducing the combinatorics of data selection to a level usable on the current generation of computing systems.

In one exemplary preferred embodiment of the present invention, a "signal-to-noise" ratio combined with a "height-and-width" ratio is used at Step 76. However, other initial thresholds can also be used, and the present invention is not limited to the initial threshold wherein described. The initial threshold is used in one exemplary preferred embodiment of the present invention as an initial threshold overview to identify a likely set of false standard biotechnology fragment peak features (e.g., in polynucleotide fragments). Data outside the initial threshold is rejected as is illustrated in FIG. 5 below. An actual sample index location of a given candidate is taken to be that of a local maximum of a peak feature, if this is unique, or alternatively to a spatial center of a feature interval.

FIG. 5 is a block diagram 86 illustrating a filtered and normalized multi-component data signal using Method 20 from FIG. 2. To illustrate the difficulty in size standard detection for polynucleotide data fragments. FIG. 5 illustrates a

relatively clean set of superficially acceptable data peaks. However, there are features 88 and 90 near sample indices 1400 and 3250, which may satisfy a signal-to-noise criterion but fail a height-and-width criterion used to determine a data peak (Items 88 and 90 of FIG. 5 correspond to items 98 and 100 of FIG. 6). The features 88 and 90 are rejected with the initial criterion at Step 76. However, there are also features 92 and 94 near sample index 2700 that meet the initial criterion, but which are *not* valid standard peaks for this exemplary biotechnology data trace (items 92 and 94 of FIG. 5 correspond to item 102 of FIG. 6). These features 92,94 are removed with the remainder of Method 74 at Steps 78-84. It is desirable to consistently remove such invalid peaks to create a valid set of standard peaks (e.g., for polynucleotide data fragments), to allow reproducible results every time an experiment is conducted.

In one exemplary preferred embodiment of the present invention, modeling physics of gel electrophoresis used to record polynucleotide data fragments is done using Fickian diffusion with drift. However, other modeling techniques could also be used and the present invention is not limited to Fickian diffusion with drift. As is known in the art, Fickian diffusion is molecular diffusion, governed by Fick's laws, which describe a rate of flow of diffusants across a unit area of a certain plane as directly proportional to a concentration gradient. For more information on Fickian diffusion see "Diffusion Processes and Their Sample Paths" by Henry P. McKean and Kiyoshi Ito, Springer Verlag, 1996, ISBN-3540606297, or "Mathematics of Diffusion" by John Crank,
Oxford University Press, 1975, ISBN-0198534116, both of which incorporated herein by reference.

Using Fickian diffusion on a gel, the drift properties of diffusants are associated with the times of arrival of their maximum concentrations at a fixed point

in a gel. For linear molecules of interest, this arrangement leads to at least three significant model predictions for polynucleotide data fragments. First, the polynucleotide data fragments drift with velocity inversely proportional to their size. Second, for sparse mixtures, fluorescence peak heights are proportional to polynucleotide data fragment counts. Finally, both of these proportionalities are independent of polynucleotide data fragment size. The value of gel electrophoresis in biomolecular size assays is due to the fact that it is possible to engineer instruments and protocols for which these predictions are valid for a significant variety of conditions and molecules.

In one exemplary preferred embodiment of the present invention, comigrating standard polynucleotide fragment sets of known size provide a means of rejecting the false peak clutter. Since an inverse proportionality between fragment size and drift velocity is independent of fragment size, and a standard fragment set is both known and ordered, a straight line drawn through a plot of standard fragment sizes as a function of their scan line locations should reveal those data peaks that are clutter. The clutter peaks will either not fall on, or sufficiently near a line, or they will cause a line to miss a significant fraction of the other data.

Given this approach to clutter rejection, there are at least two remaining problems in applying it to biotechnology data. First, potential combinatorics of quickly choosing an appropriate subset of valid peaks from candidate peaks can be computationally impossible or forbidding for currently available computing systems. Secondly, a degree to which an inverse proportionality of fragment and drift velocity size is genuinely independent of fragment size depends upon a degree to which gel properties are consistent and uniform over a period of observation.

FIG. 6 is a block diagram 96 illustrating filtered standard polynucleotide fluorescence responses for a sequence of scans for a set of lanes in a gel which were loaded with standard polynucleotide fragments at a same time. The physical edges of the gel correspond to the edges of this image, and the bright bands in any one lane represent the scan line locations of candidate standard fragments in that lane. For example, the three scan lines near sample index 2000 (FIG. 6) represent the three data peaks near sample index 2000 (FIG. 5). Note the smaller bright features 98, 100 and 102, roughly in the center of lanes 10, 19, and 25, that do not belong to bands that extend across the image. These are examples of the "false peak clutter" at issue. For example, item 98 (FIG. 6) may correspond to false peak 88 (FIG. 5), item 100 may correspond to false peak 90 (FIG. 6) and item 102 (FIG. 6) may correspond to false peaks 92,94 (FIG. 5).

If the properties of the gel were uniform throughout the gel over a period of successive scans, the bright bands would be strictly horizontal (e.g., exemplary horizontal dashed line 104). Not only are the bands *not* horizontal, the degree to which they curve increases as a function of time, with larger scan lines indices corresponding to scans occurring later in time. The drifting fragments in the gel are charged particles moving through a resistive medium under the influence of an applied electric field. The resulting characteristic "smile" (e.g., scan line 106 versus horizontal line 104) in such electrophoretic gel imagery is due to the differential heating of the gel by this current over time, the edges of the gel more effectively dissipating heat than the more central regions.

The smaller a linearly ordered set of standard fragment sizes (*e.g., a mask*) is, the more the resulting combinatorics of selecting a valid subset (*e.g., flickering a mask*) become tractable. The more localized overlapping regions of the gel to which

each mask is applied, the more uniform and consistent the relevant gel properties become.

In one exemplary preferred embodiment of the present invention, a given a set of candidate standard peak scan line locations are obtained at Step 76 by the initial threshold criterion outlined above. In such an embodiment, clutter and false peak rejection proceeds by choosing proper, overlapping subsets of a complete standard size set at Step 78.

At Step 78, linear mappings are applied to the multiple overlapping subsets of data points. For an ordered, sequential three element set of standard sizes $\{ M_a, M_b, M_c \}$ whose peaks occur at scan lines $\{ n_a, n_b, n_c \}$, respectively, linear regression techniques give a predictive linear mapping of scan line n_x to fragment size as is illustrated in Equation 5. However, other set sizes and linear mappings could also be used and the present invention is not limited to the linear mappings in Equation 5.

$$\mu^{(0)}_{abc} + \mu^{(1)}_{abc} * n_x \qquad (5)$$

The coefficients $\{ \mu^{(j)}_{abc} \}$ are functions of a particular set of (size, scan line) pairs. With any scan line n lying between two consecutive standard peak scan line locations, $\{ n_b, n_c \}$, a local Southern linear mapping method associates a fragment size as is illustrated in Equation 6. However, other linear mapping methods can also be used, and the present invention is not limited to the local Southern method linear mappings illustrated in Equation 6.

$$M'_n \equiv (\mu^{(0)}_{abc} + \mu^{(1)}_{abc} * n + \mu^{(0)}_{bcd} + \mu^{(1)}_{bcd} * n) / 2 \qquad (6)$$

The set $\{M_b, M_c, M_d\}$ is a rightmost overlapping "bcd" and sequential set of standard sizes for a leftmost overlapping "abc" and sequential set $\{ M_a, M_b, M_c \}$, the former for standard size peaks occurring at scan lines $\{ n_b, n_c, n_d \}$. An individual error in

this association of standard peak size (i.e., data point value) and scan line location (i.e., data point) is calculated as a difference illustrated by Equation 7.

$$\varepsilon_n \equiv M_n - M'_n \qquad (7)$$

At Step 82, multiple error values (e.g., Equation 7) are determined from the application of multiple linear mappings (e.g., Equation 6) to the multiple overlapping subset of data points. In one preferred embodiment of the present invention, a Root Mean Square ("RMS") error evaluation of the "goodness" of each of the local fits allows them to be ranked. However, other error evaluation methods can also be used and the present invention is not limited to RMS.

Given a set of peak scan line locations for a set of standard biotechnology fragments sizes, straight lines are fit to possible sets of three adjacent fragment sizes as a function of the three associated adjacent scan line locations, using linear regression. A local linear mapping of any given scan line to its associated fragment size is then formed by averaging the two most relevant of these three-point linear fits.

A first relevant fit includes two closest standard scan lines, which are smaller than a given scan line, and one closest standard scan line, which is greater. A second relevant fit includes two closest standard scan lines, which are greater than a given scan line, and one closest standard scan line which is smaller. A total RMS error over the **K** (size, scan line) pairs $\{ (M_{n(k)}, n(k)) \}$ is illustrated in Equation 8.

$$\text{error} = [\textstyle\sum_{k=1,\ldots,K} \varepsilon^2_{n(k)} / K]^{1/2} = [\textstyle\sum_{k=1,\ldots,K} (M_{n(k)} - M'_{n(k)})^2 / K]^{1/2} \qquad (8)$$

A set of subsets of scan line locations which yields a smallest total RMS error is chosen at Step 84, provided that both a total error and an error for any one standard size are below certain error thresholds. If these error thresholds cannot be satisfied by

any subset of scan line locations for a complete set of standard sizes, a size of a standard size set is reduced by one and the error calculation is repeated. This method of evaluating local linear fits to possible subsets of standard scan line locations is repeated, over possible standard size sets of the reduced size. The RMS process (e.g., Equation 8) is repeated until either error threshold criterion are satisfied, or until a reduced size of the standard size set becomes too small. There is also a selection criterion on the subsets of the complete standard size set that prevents more than a given number of adjacent lacunae in a final size set.

FIG. 7 is a block diagram 108 illustrating exemplary biotechnology peaks (e.g., polynucleotide peaks) using size standard detection with false peak clutter rejection from Method 74 of FIG. 4. Target biotechnology fragment peaks 110, 112, 114, 116, 118, 120, 122, 124, 126 and 128 identified by Method 80 (FIG. 4) while standard biotechnology peaks (e.g., sample indices for known polynucleotide data sequences) are indicated by dashed vertical lines. For example, the dashed line through the data peak 110 indicates a known polynucleotide fluorescence intensity. The false peaks 88,90 (FIG. 5) near scan lines 1400 and 3250 that may satisfy a signal-to-noise criterion but fail a height-and-width criterion are properly identified and removed with initial criterion at Step 76 of Method 80. The false peaks 92,94 (FIG. 5) have been properly identified and rejected as clutter by the remaining steps of Method 80. Note that several of the data peaks (e.g., 114, 118, 122) for target data do no line up exactly on a dashed line for known data. Such data peaks are adjusted as is described below.

Method 74 (FIG. 4) may also allow for the application of a number of very powerful and convenient quality control measures. First, Method 74 may implicitly bootstrap a sizing calibration. This allows a quality of fluorescence intensity data to

be immediately assessed from their susceptibility to accurate calibration. This may be an effective measure of the degree of conformance between experimental data and a good physical model of the processes implicated in their creation. Secondly, limits are placed on both the total number and distribution of size standards fragments that can be deleted from the initial set in producing a set of local linear mappings with acceptable error. Finally, it is assumed that false peak clutter usually has its source in either residual spectral bleed-through, or more problematically for any given lane, standard fragment sets which actually belong to adjacent lanes. This latter phenomenon is known as "cross-talk." By keeping track of both how many candidate standard peak scan line locations co-occur in adjacent lanes as well as how many detected standard peaks are co-located in adjacent lanes even after application Method 74, it is possible to form yet another useful data quality measure. This measure may be particularly relevant to clutter rejection because it essentially qualifies its self-consistency.

Data size calibration and adjustment

The actual size and location of the filtered and false peak clutter rejected data (e.g., polynucleotide fragment output) is typically adjusted to allow experimental data to be more accurately visually displayed. This adjustment provides more accurate data values for visual display. For example, target data peaks illustrated in FIG. 7 that do not line up exactly on a known data peak values are adjusted.

FIG. 8 is a block diagram illustrating a Method 130 for data size calibration and adjustment. At Step 132, a first final subset of overlapping data points with a smallest error value is selected as a standard set of data points from a first set of data points. Data points in the first final subset of overlapping data points include data points with values that fall within a standardized range and where false data points

have been removed. At Step 134, higher order mappings are applied to the first final subset of data points to further reduce the smallest error value for the final subset of overlapping data points and create a second final subset of data points.

In one preferred embodiment of the present invention, a first subset of overlapping data points is selected at Step 132 from application of Method 74 (FIG. 4). However, other methods can also be used to select the final subset of overlapping data points, and the present invention is not limited to the application of Method 74.

At Step 132, the first final subset of overlapping data points selected from application of Method 74 including a local Southern method (e.g., Equations 5 and 6), size-calibrates data with a limited precision (e.g. typically no better than one to two base pairs for polynucleotide fragment data). If the data points can be calibrated in Step 132 to within a pre-determined quality control limit, the local Southern calibration is followed by a higher order mapping at Step 134 that further reduces a calibration error. In one exemplary preferred embodiment of the present invention, the calibration error is reduced to zero. In another exemplary preferred embodiment of the present invention, the calibration error is reduced to a very small value approaching zero, but not to zero (i.e., slightly greater than zero).

Method 130 combines the local statistical robustness of regression techniques (i.e., with their natural rejection of outliers) and a precision possible with higher order methods (e.g., higher order splines). In one exemplary preferred embodiment of the present invention, absolute precision in the calibration biotechnology data is desired to provide accurate and reproducible results. However, the present invention can also be used if only relative precision is desired.

At Step 134, higher order mappings are used with the residual error from the local Southern Method, and a second-order generalization of that linear, or first-order

local Southern Method. In one exemplary preferred embodiment of the present invention, local quadratic or second-order maps are constructed using residual errors for the same three element sets of (fragment size, scan line location) pairs used for the Local Southern Method. However, the present invention is not limited to second order maps and higher order maps can also be used (e.g., third order, fourth order, etc.).

Since a second-order mapping has three coefficients, or three "degrees of freedom," the three residual errors for each set of three pairs can, in principal, be accounted for in a very exact manner. Control of computational degeneracy in a numerical order of an error is accomplished by using a singular value decomposition to solve a linear system of equations that a conventional least squares method produces when fitting a quadratic to three data points.

Given the local Southern approximation of a size associated with any specific scan line location, an additive correction higher order mapping is formed by averaging two most relevant of these second three-point quadratic fits. A first approximation, for two closest standard scan lines which are smaller than a given scan line and one closest standard scan line which is greater. A second approximation for two closest standard scan lines which are greater than a given scan line and one closest standard scan line which is smaller. Since each quadratic fit is locally exact at the scan line locations of relevant three standard fragment peaks, averaging any two fits on these peak locations is also exact, which results in an absolutely precise interpolation on the detected standard fragment set.

For a scan line **n**, the local Southern method (e.g., Equations 5 and 6) associates a fragment size $\mathbf{M'}_n$, with error ε_n at the standard peak locations. With the same notation and conventions used for the discussion of the local Southern method

above, a least squares method gives exact second order mappings of an error at any one standard peak location for leftmost sequential set of standard sizes as illustrated in Equation 9. However, other methods can also be used and the present invention is not limited to a least squares methods.

$$\gamma^{(0)}_{abc} + \gamma^{(1)}_{abc} * n + \gamma^{(2)}_{abc} * n^2 \qquad (9)$$

Exact second order mappings of an error at any one standard peak location for rightmost sequential set of standard sizes is illustrated in Equation 10.

$$\gamma^{(0)}_{bcd} + \gamma^{(1)}_{bcd} * n + \gamma^{(2)}_{bcd} * n^2 \qquad (10)$$

Both sets of coefficients $\{ \gamma^{(j)}_{abc} \}$ and $\{ \gamma^{(j)}_{bcd} \}$ are functions of their respective particular set of (size, scan lines) pairs and the error ε_n. For any scan line n lying between two consecutive standard peak scan line locations, $\{ n_b, n_c \}$, a higher-order residual mapping adds a correction factor δ_n to a local Southern method size association as illustrated in Equation 11.

$$\delta_n \equiv (\gamma^{(0)}_{abc} + \gamma^{(1)}_{abc} * n + \gamma^{(2)}_{abc} * n^2 + \gamma^{(0)}_{bcd} + \gamma^{(1)}_{bcd} * n + \gamma^{(2)}_{bcd} * n^2) /2 \qquad (11)$$

In one preferred embodiment of the present invention, this correction δ_n, or higher order mapping, gives a net association that is exact at scan line locations of the standard peak features. However, the present invention is not limited to such a correction δ_n and other correction features could also be used.

FIGS. 9A and 9B are block diagrams 136, 138 illustrating data size calibration using Method 130 from FIG. 8. FIG. 9A illustrates an exemplary data peak 140 (e.g., for an unknown polynucleotide sequence) before application of Method 130 (FIG. 8). The data peak 140 is slightly offset from a relevant desired data peak location 142 (e.g., for a known polynucleotide sequence) whose desired location is illustrated by a

dashed line, that would be achieved if there were no errors for a data set acquired from a desired experiment. FIG. 9B illustrates an exemplary data peak 144 after application of Method 130 (FIG. 8). The data peak 146 is more accurately aligned over the desired data peak location 142 after application of Method 130.

FIGS. 9A and 9B illustrates only one exemplary data peak. However, Method 130 is applied to all data peaks (e.g., 54, 56, 58, 60, 62, 64, 66, 68, 70 and 72 of FIG. 3D) in a final subset of overlapping data points (e.g., produced by Method 74 of FIG. 4) to further reduce error for a set of data points that will be visually displayed. Method 130 may improve a set of data points that will be displayed and analyzed by further reducing data errors that may be introduced as a result of running a desired experiment.

Data peaks that have been sized and adjusted may still include data "stutter." (See e.g., FIG. 11A). For example, the data peaks illustrated in the figures are illustrated as a "smooth" data peaks. However, actual experimental data peaks typically include multiple sub-peaks, that are a function of the actual data. It is desirable to remove the multiple sub-peaks, or data stutter before visual display.

Reduction of data magnitude and data smoothing

In the current generation of biotechnology equipment known in the art, scan lines from gel-electrophoresis are formed at a rate which, after size calibration, results in an over-resolution of the sized traces by about an order of magnitude. That is, there are about ten scan lines between each successive integer base-pair value. In addition, biotechnology fragments (e.g., polynucleotide fragments) typically occur in cluster around the most significant fragment sizes, rather than as cleanly isolated peaks of integer base-pair width. This can be seen by comparing the broader and more complex peak features (e.g., feature 44) in the biotechnology fragment trace in Figure

3C, with the narrow and more simple standard fragment peaks in Figure 3D (e.g., data point 68).

Representing these complex biotechnology fragment traces at their full resolution on the windowed display 16 is further complicated by the inevitable limits imposed by the current generation computer monitor and graphics display systems. Consequently, before creating graphical images to display, the biotechnology data points are further decimated and smoothed using an "envelope detector" that enhances a visibility of data points for display on the windowed display 16 by moderating resulting fragment "stutter."

FIG. 10 is a flow diagram illustrating a Method 146 for envelope detection. At Step 148, an envelope criterion is established for sub-sampling of a second final subset of overlapping data created from a first final subset of overlapping data. The second final subset of overlapping data points have been adjusted to fall within a standard size. Significant features of the second final subset of overlapping data are preserved within the envelope criterion. At Step 150, the envelope criterion is applied to compress the number of data values in the second final subset of overlapping data by at least one order of magnitude, reduce data stutter, and to create a third final subset of overlapping data.

In one exemplary preferred embodiment of the present invention, the second final subset of overlapping data is produced by applying Method 20 (FIG. 2), Method 74 (FIG. 4) and Method 130 (FIG. 8) discussed above. However, the present invention is not limited to overlapping data sets produced with these method and other data sets produced with other methods known in the art, that will be displayed on the windowed display 16 can also be used with Method 146 (FIG. 9).

In one exemplary preferred embodiment of the present invention, the envelope criterion established at Step 148 is based on a "nonlinear box-car-extremum" filter that compresses data size resolution by about an order of magnitude and removes data stutter. However, other envelope criterion could also be used and the present invention is not limited to a nonlinear box-car-extremum filter.

In one preferred embodiment of the present invention, graphical images for the windowed display 16 illustrate a size resolution of about one polynucleotide base pair, with each point on a trace sampled at integer base-pair sizes. At Step 150, the box-car envelope detector first segments a size axis of a size-calibrated full resolution trace data into contiguous regions centered on these integer sizes. The term "box-car" reflects the view of these contiguous, disjoint regions as box-cars aligned end-to-end along a size axis.

A trace envelope is formed by replacing signal intensities associated with sizes in a given box-car by their maximum. This is a many-to-one replacement, or "decimation", on the order of the average number of scan lines associated with an integer base pair in the full resolution data. Preferably, this decimation factor is about ten-to-one. However, other decimation factors can also used.

In one exemplary preferred embodiment of the present invention, at Step 150, an envelope criterion $\mathbf{f}^*_k$, is applied in Equation 12.

$$\mathbf{f}^*_k \equiv \max \{ \mathbf{f}_0(n) : (\mathbf{M}^*_k + \mathbf{M}^*_{k-1}) /2 \leq (\mathbf{M}'_n + \delta_n) < (\mathbf{M}^*_{k+1} + \mathbf{M}^*_k) /2 \} \quad \textbf{(12)}$$

The notation and conventions in Equation 12 reflect notation from Equations 1-11 discussed above. For example, $\mathbf{f}_0$ is determined with Equation 4, $\mathbf{M}'_n$ with Equation 6, and δ_n with Equation 11, etc.

FIGS. 11A and 11B are block diagrams 152,154 illustrating envelope detection using Method 146 of FIG. 10. FIG. 11A illustrates an envelope 156 created around a target data peak 158. Data "stutter" is illustrated by two small peaks on the left side (i.e., towards 2000 sample index), and one small peak on the right side (i.e., towards 2500 sample index) of target data peak 158. FIG. 11B illustrates a new data peak 160 after application of Method 146. The number of data points in the new data peak 160 is reduced by an order of magnitude and the "stutter" of the data peak 158 has been removed. FIGS. 11A and 11B illustrates only one exemplary data peak. However, Method 150 is applied to data peaks in the second final subset of overlapping data. Data peaks described herein, also typically include data "stutter." However, data peaks in other than FIG. 11A are illustrated as smooth and do not illustrate data stutter that does exist before application of Method 146 simplify the drawing of such data peaks.

Method 146 may further enhance a visibility of data points for display on the windowed display 16 by moderating resulting fragment "stutter." The number of data points may also be reduced by an appropriate amount (e.g., one order of magnitude) for easier display.

Processing of general multi-component signal data

In one exemplary preferred embodiment of the present invention, a general multi-component data signal can be processed to yield a set of data peaks for a target experiment suitable for display on the windowed display 16 of the display device 14. In such an embodiment, the general multi-component data signals may include general biotechnology multi-component data signals. However, the present invention is not limited to processing general biotechnology multi-component signal data, and other signal data could also be processed (telecommunications signals, electrical

signals data for electrical devices, optical signals, physical signals, or other data signals).

FIGS. 12A and 12B is a flow diagram illustrating a Method 162 for processing experimental data. At Step 164, of FIG. 12A, a multi-component data signal is read. The multi-component data signal includes multiple individual data signal components of varying spectral characteristics and varying amplitudes. The multiple individual data signal components overlap within portions of the multi-component data signal. At Step 166, filters are applied to the multi-component data signal to create multiple non-overlapping individual data signal components. The filter also filters multiple signal artifacts in the multi-component data signal that introduce ambiguity to base values in the multiple non-overlapping individual data signal components to spatially detrend and normalize the multiple non-overlapping individual data signal components to a uniform set of base values. At Step 168, multiple linear mappings are applied to multiple overlapping subsets of data points from the multiple non-overlapping individual data signal components to select a first final subset of overlapping data points with a smallest error value. The data points in the first final subset of overlapping data points include data points that fall within a standardized range and wherein false data points have been removed.

At Step 170 of FIG. 12B, multiple higher order mappings are applied to the first final subset of overlapping data points to further reduce the smallest error value for the final subset of overlapping data points and create a second final subset of data points. At Step 172, an envelope criterion is applied to compress the number of data values in the second final subset of overlapping data by at least an order of magnitude, reduce data stutter, and create a third final subset of overlapping data. Significant features of the second final subset of overlapping data are preserved within the

envelope criterion. The third final subset of overlapping data is suitable for the windowed display 16 on the display device 14.

Method 162 allows the processing of multi-component data signals from biotechnology experiments or experiments from other arts to be automated. When a multi-component data signal is input, a third final subset of overlapping data with multiple data peaks suitable for display on a windowed device is automatically produced. This may help reduce or eliminate inconsistencies in experimental data processing that typically lead to unreliable or erroneous results.

In one exemplary preferred embodiment of the present invention, the multi-component data signal includes multi-component fluorescence intensities for polynucleotide data including DNA, cDNA or mRNA. However, the present invention is not limited to multiple-component data signals for polynucleotide data, or other biotechnology data, and multi-component data signals from other arts can also be used (e.g., telecommunications signals, electrical signals data for electrical devices, optical signals, physical signals, or other data signals).

In yet another exemplary preferred embodiment of the present invention, Method 162 is accomplished by applying Method 20 (FIG. 2) at Steps 164, 166 (FIG. 12A), Method 74 (FIG. 4) at Step 168 (FIG. 12A), Method 130 (FIG. 8) at Step 170 (FIG. 12B), and Method 146 (FIG. 10) at step 172 (FIG. 12B). However, the present invention is not limited to applying all the steps of these methods to accomplished Method 162 (FIGS. 12A and 12B). Method 162 can be accomplished by applying selected steps from these methods.

FIGS. 13A and 13B are block diagrams 174, 176 illustrating Method 162 of FIGS. 12A and 12B. FIG. 13A illustrates a multi-component data signal 178 of interest. FIG. 13B illustrates set of processed desired data peaks 180, 182, 184, 186.

188, 190, 192, 194, 196, 198, 200 from the multi-component data signal 178 after processing with Method 162. The multi-component data signal has been filtered, normalized to a predetermined size, had false peaks, errors and data stutter removed, has been smoothed, and had the number of data values reduced by at least one order of magnitude. The processed desired data peaks are suitable for display on the windowed display 16 of the display device 14.

In one exemplary preferred embodiment of the present invention, the desired data peaks 180, 182, 184, 186, 188, 190, 192, 194, 196, 198 and 200 (FIG. 13B) are polynucleotide fragment peaks (e.g., DNA, cDNA or mRNA). However, the present invention in not limited to multi-component data signals including polynucleotide fragment data and other multi-component data signals including other experimental information could also be used (e.g., telecommunications signals, electrical signals data for electrical devices, optical signals, physical signals, or other data signals).

Exemplary multi-component data processing system

FIG. 14 is a block diagram illustrating an exemplary multi-component data processing system 202. The multi-component data processing system includes a data sample and reference calibration module 204, an optional broadband signal collection module 206, a storage module 208, a filtering and baseline module 210, a reference and sample calibration module 212 and a display module 214.

The data sample and reference calibration module 204 is used for processing known and target biotechnology samples. The optional broadband signal collection module 206 is used for collecting experimental data from multi-component data signals when laser-induced fluorescence of biotechnology products is used. In another embodiment of the present invention, the optional broadband signal collection module 206 can be eliminated if other technologies are used instead of laser-induced

fluorescence (e.g., micro-arrays). The storage module 208 is used to store experimental data. The filtering and baseline module 210 is used to remove spectral overlap and normalize experimental data if laser-induced fluorescence is used, or can be used to perform other filtering and baselines if other technologies are used (e.g., micro-arrays).

The reference and calibration module 212 is used for standard size detection with false peak and clutter removal, data size calibration, envelope detection and data stutter removal of experimental data. The display module 214 visual displays processed experimental data. However, the present invention is not limited to these modules and more or fewer modules could also be used. In additional, the functionality of the modules described could be combined or split into additional modules.

In one exemplary preferred embodiment of the present invention, experimental data processing system 10 (FIG. 1) includes the storage module 208, the filtering and baseline module 210, the reference and sample calibration module 212 and the display module 214 (FIG. 14) as an integral combination of hardware and software (i.e., indicated by the dashed line in FIG. 14). This allows virtually any experimental technique (e.g., gel-electrophoresis, miro-arrays, etc.) to be used to generate data files that are stored in the storage module 208 and processed with the methods described herein with software resident on the computer 12. Such an embodiment provides flexibility to process experimental data from a wide variety of applications on a conventional personal computer system, or other larger computer system.

The methods and system described herein are used to process data for display on the windowed display 16 of display device 14, as is illustrated by FIG. 13B. However, the final set of data (e.g., the third final subset of data) may still require

additional processing to aid in an analysis of displayed data. As was discussed above, one of the most commonly used methodologies in biotechnology is comparison. Comparisons of experimental data to known data presents some additional problems such as experiment-to-experiment variability that are overcome with methods and system described in co-pending Application No. 09/318,679, assigned to the same Assignee as the present application.

The methods and system described herein help automate the processing of experimental data to eliminate or reduce errors and leave processed experimental data in a format suitable for visual display. The methods and systems may help reduce or eliminate inconsistencies in experimental data processing that typically lead to unreliable or erroneous results.

It should be understood that the programs, processes, methods and system described herein are not related or limited to any particular type of computer or network system (hardware or software), unless indicated otherwise. Various types of general purpose or specialized computer systems may be used with or perform operations in accordance with the teachings described herein.

In view of the wide variety of embodiments to which the principles of the present invention can be applied, it should be understood that the illustrated embodiments are exemplary only, and should not be taken as limiting the scope of the present invention. For example, the steps of the flow diagrams may be taken in sequences other than those described, and more or fewer elements may be used in the block diagrams. While various elements of the preferred embodiments have been described as being implemented in software, in other embodiments hardware implementations may alternatively be used and visa-versa.

The claims should not be read as limited to the described order or elements unless stated to that effect. Therefore, all embodiments that come within the scope and spirit of the following claims and equivalents thereto are claimed as the invention.

WE CLAIM:

1. A method for normalizing data, comprising the following steps:

reading a multi-component data signal, wherein the multi-component data signal includes a plurality of individual data components of varying spectral characteristics and varying amplitudes, and wherein the plurality of individual data signal components overlap within portions of the multi-component data signal;

applying a spectral filter to the multi-component data signal to create a plurality of non-overlapping individual data signal components: and

applying a spatial filter to a plurality of signal artifacts in the multi-component data signal that introduce ambiguity to base values in the plurality of non-overlapping individual data signal components to spatially detrend and normalize the plurality of non-overlapping individual data signal components to a uniform base value.

2. A computer readable medium having stored therein instructions for causing a central processing unit to execute the method of Claim 1.

3. The method of Claim 1 wherein the step of applying a spectral filter includes applying a demultiplexing filter.

4. The method of Claim 3 wherein the demultiplexing filter includes applying:

$$A'(p) = \Sigma_q m(p,q)\, A(q),$$

wherein A'(p) is an unfiltered measured fluorescence intensity , m(p,q) is a coefficient matrix for a measurement of an amount of energy measured at a wavelength that corresponds a p-th data point in a center of a fluorescence response for a q-th data point, and A(q) is a measurement of an actual fluorescence intensity, and wherein A(q) is determined by inverting a resulting linear system of equations using a singular value decomposition of a coefficient matrix m(p,q).

5. The method of Claim 1 wherein the step of applying a spatial filter includes applying a nonlinear morphological gray-scale rolling-ball transformation filter.

6. The method of Claim 5 where applying the nonlinear morphological gray-scale rolling-ball transformation filter includes applying:

$$f_-(n) \equiv \min \{ f(n+m) - f_{min} : -N/2 \le m \le N/2 \};$$
$$f_\pm(n) \equiv \max \{ f_-(n+m) + f_{min} : -N/2 \le m \le N/2 \}; \text{ and}$$
$$f_0(n) \equiv f_-(n) - f_\pm(n);$$

wherein, $f_-(n)$ is an erosion of a fluorescence intensity f(n) measured at a data point-n from a set of N-data points including data point n, $f_\pm(n)$ is a dilation of $f_-(n)$, and $f_0(n)$ is a fluorescence intensity of a rolling ball filtered version of a data point-n, and wherein finding $f_-(n)$ and $f_\pm(n)$ results in a nonlinearity.

7. The method of Claim 1 wherein the multi-component data signal includes polynucleotide data.

8. The method of Claim 7 wherein the polynucleotide data includes any of DNA. cDNA or mRNA data.

9. A method for clutter rejection. comprising the following steps:

selecting a first set of data points from a filtered set of data points using initial threshold criterion;

selecting a plurality of overlapping subsets of data points from the first set of data points;

applying a plurality of linear mappings to the plurality of overlapping subset of data points;

determining a plurality of error values from the applying of the plurality of linear mappings to the plurality of overlapping subset of data points; and

selecting a first final subset of overlapping data points with a smallest error value from the first set data points, wherein data points in the first final subset of overlapping data points include data points that fall within a standardized range and wherein false data points have been removed.

10. A computer readable medium having stored therein instructions for causing a central processing unit to execute the method of Claim 9.

11. The method of Claim 9 wherein the step of selecting a plurality of data points using an initial threshold criterion includes applying a signal-to-noise ratio and a height-and-width ratio to the filtered set of data points.

12. The method of Claim 9 wherein the step of selecting a first set of data points from a filtered set of data points using initial threshold criterion includes:

applying a spectral filter to a multi-component data signal to create a plurality of non-overlapping individual data signal components:

applying a spatial filter to a plurality of signal artifacts in the multi-component data signal that introduces ambiguity to base values in the plurality of non-overlapping individual data signal components to spatially detrend and normalize the plurality of non-overlapping individual data signal components to a uniform base value and to create a filtered set of data points; and

selecting a first set of data points from the filtered set of data points using a signal-to-noise ratio and a height-and-width ratio.

13. The method of Claim 9 wherein the step of selecting a plurality of overlapping subsets of data points from the first set of data points includes selecting sequential three element sets $\{ M_a, M_b, M_c\}$ whose maximum values occur at data points $\{n_a, n_b, n_c\}$.

14. The method of Claim 9 wherein the step of applying a plurality of linear mappings to the plurality of overlapping subsets of data points includes applying a plurality of local Southern linear mappings to the plurality of overlapping subset of data points.

15. The method of Claim 14 wherein the local Southern linear mapping includes applying:

$$M'_n \equiv (\mu^{(0)}_{abc} + \mu^{(1)}_{abc} * n + \mu^{(0)}_{bcd} + \mu^{(1)}_{bcd} * n) /2,$$

wherein M'_n is a predictive linear mapping of a data point-n to a value for the data point-n and wherein "abc" in coefficients $\mu^{(0)}_{abc}$, $\mu^{(1)}_{abc}$, indicates a leftmost overlapping subset of data points, and "bcd" in coefficients $\mu^{(0)}_{bcd}$, and $\mu^{(1)}_{bcd}$, indicates is a rightmost overlapping subset of datapoints.

16. The method of Claim 9 wherein the step of determining a plurality of error values from the applying of the linear mappings to the plurality of overlapping subset of data points includes determining a plurality of error values using a Root Mean Square error evaluation of the linear mappings.

17. The method of Claim 16 wherein the Root Mean Square error evaluation includes applying:

$$\text{Error} = [\ \Sigma_{k=1,\ldots,K}\ (\ M_{n(k)} - M'_{n(k)}\ )^2 / K\ ]^{1/2},$$

wherein Error is a total Root Mean Square error over K-pairs of sets of data points, $M_{n(k)}$ is an actual value of a data point-n, and $M'_{n(k)}$ is a predictive linear mapping of a data point-n to a value for a data point-n over K-pairs of sets of data points.

18. The method of Claim 9 wherein the step of selecting a first final subset of overlapping data points with a smallest error value as a standard set of data points from the first set data points includes selecting a smallest error value ε_n by applying:

$$\varepsilon_n \equiv M_n - M'_n$$

wherein M_n is an actual value of a data point-n. and M'_n is a predictive linear mapping of the data point-n to a value for the data point-n.

19. The method of Claim 9 wherein the first set of data points includes polynucleotide data.

20. The method of Claim19 wherein the polynucleotide data includes any of DNA, cDNA or mRNA data.

21. A method for data size calibration. comprising the following steps:

selecting a first final subset of overlapping data points with a smallest error value from a first set data points, wherein data points in the first final subset of overlapping data points include data points that fall within a standardized range and wherein false data points have been removed; and

applying a plurality of higher order mappings to the first final subset of data points to further reduce the smallest error value for the first final subset of overlapping data points and create a second final subset of data points.

22. A computer readable medium having stored therein instructions for causing a central processing unit to execute the method of claim 21.

23. The method of Claim 21 wherein the step of selecting a first final subset of overlapping data points with a smallest error value includes selecting a first final subset of overlapping data points by applying a plurality of linear mappings to a plurality of overlapping subsets of data points and determining a plurality of error

values derived from applying the plurality of linear mappings to the plurality of overlapping subset of data points.

24. The method of Claim 21 wherein the step of determining a plurality of error values includes determining a plurality of error values using a Root Mean Square errors of the linear mappings.

25. The method of Claim 21 wherein the step of applying a plurality higher order mappings to the first final subset of data points to further reduce the smallest error value for the final subset of overlapping data points includes reducing the smallest error value to a value greater than zero.

26. The method of Claim 21 wherein the step of applying a plurality of higher order mappings to the first final subset of data points to further reduce the smallest error value for the final subset of overlapping data points includes reducing the smallest error value to zero.

27. The method of Claim 21 wherein the step of applying a plurality of higher order mappings to the first final subset of data points to further reduce the smallest error value for the final subset of overlapping data points includes applying a higher order residual snapping method to further reduce residual error from a plurality of linear mappings used to create the final subset of overlapping data points.

28. The method of Claim 27 wherein the residual error is reduced to zero.

29. The method of Claim 21 wherein a higher order mapping includes applying a second order mapping.

30. The method of Claim 21 wherein the step of applying a plurality of higher order mappings includes applying:

$$\delta_n \equiv (\gamma^{(0)}_{abc} + \gamma^{(1)}_{abc} * n + \gamma^{(2)}_{abc} * n^2 + \gamma^{(0)}_{bcd} + \gamma^{(1)}_{bcd} * n + \gamma^{(2)}_{bcd} * n^2) /2,$$

wherein δ_n is a correction factor providing a second order mapping from a linear mapping of a data point-n to a value for the data point-n and wherein "abc" in coefficients $\gamma^{(0)}_{abc}$, $\gamma^{(1)}_{abc}$, and $\gamma^{(2)}_{abc}$ is a leftmost overlapping subset of data points, wherein "bcd" in coefficients $\gamma^{(0)}_{bcd}$, $\gamma^{(1)}_{bcd}$, and $\gamma^{(2)}_{bcd}$ is a rightmost overlapping subset of data points, and wherein the coefficients $^{0)}_{abc}$, $\gamma^{(1)}_{abc}$, $\gamma^{(2)}_{abc}$, $\gamma^{(0)}_{bcd}$, $\gamma^{(1)}_{bcd}$, $\gamma^{(2)}_{bcd}$ are functions of their respective subset of data points and an error value for a linear mapping.

31. The method of Claim 21 wherein the first final subset of overlapping data points includes polynucleotide data.

32. The method of Claim 32 wherein the polynucleotide data includes any of DNA, cDNA or mRNA data.

33. A method for envelope detection, comprising the following steps:

establishing an envelope criterion for sub-sampling of a second final subset of overlapping data created from a first final subset of overlapping data, wherein the second final subset of overlapping data points have been adjusted to fall within a

standard size calibration, and wherein significant features of the second final subset of overlapping data are preserved within the envelope criterion; and

applying the envelope criterion to compress the number of data values in the second final subset of overlapping data by at least an order of magnitude, reduce data stutter, and create a third final subset of overlapping data.

34. A computer readable medium having stored therein instructions for causing a central processing unit to execute the method of Claim 33.

35. The method of Claim 33 wherein the step of establishing an envelope criterion includes establishing a box-car filter for the second final subset of overlapping data.

36. The method of Claim 33 wherein the step of establishing an envelope criterion includes dividing the second final subset of data into contiguous regions centered on pre-determined data points, wherein the pre-determined data points include a sizing on one or more display axis for a display device that will be used to display processed data from the second final subset of overlapping data.

37. The method of Claim 33 wherein the establishing step includes identifying macroscopic properties of the second final subset of overlapping data.

38. The method of Claim 33 wherein the applying step includes replacing a selected number of data points in the second final subset of overlapping data with a

smaller number of data points than the selected number of data points determined by the envelope criterion.

39. The method of Claim 33 wherein the applying step includes replacing sets of ten data points from the second final subset of overlapping data with one maximum data point from a subset of the second final subset of overlapping data determined by the envelope criterion.

40. The method of Claim 33 wherein step of applying the envelope criterion includes applying:

$$f^*_k \equiv \max \{ f_0(n) : (M^*_k + M^*_{k-1}) /2 \leq (M'_n + \delta_n) < (M^*_{k-1} + M^*_k) /2 \},$$

wherein f^*_k is an envelope criterion, M^*_k is an integral size sample interval used to sample the second final subset of overlapping data, $f_0(n)$ is a set of spectrally and spatially filtered fluorescence data points-n, M'_n is a predictive linear mapping of a data point-n to a value for the data point-n, and δ_n is a higher order mapping to further reduce a smallest error value from the predictive linear mapping M'_n.

41. The method of Claim 33 wherein the second final subset of overlapping data points includes polynucleotide data.

42. The method of Claim 41 wherein the polynucleotide data includes any of DNA, cDNA or mRNA data.

43. A method for processing multi-component signal data, comprising the following steps:

reading a multi-component data signal, wherein the multi-component data signal includes a plurality of individual data signal components of varying spectral characteristics and varying amplitudes, and wherein the plurality of individual data signal components overlap within portions of the multi-component data signal;

applying filters to the multi-component data signal to create a plurality of non-overlapping individual data signal components and to remove a plurality of signal artifacts in the multi-component data signal that introduce ambiguity to base values in the plurality of non-overlapping individual data signal components to spatially detrend and normalize the plurality of non-overlapping individual data signal components to a uniform base value;

applying a plurality of linear mappings to a plurality of overlapping subsets of data points from the plurality of non-overlapping individual data signal components to select a first final subset of overlapping data points with a smallest error value, wherein data points in the first final subset of overlapping data points include data points that fall within a standardized range and wherein false data points have been removed;

applying a plurality of higher order mappings to the first final subset of overlapping data points to further reduce the smallest error value for the first final subset of overlapping data points and create a second final subset of data points; and

applying an envelope criterion to compress the number of data values in the second final subset of overlapping data by at least an order of magnitude, reduce data stutter, and create a third final subset of overlapping data wherein significant features of the second final subset of overlapping data are preserved within the envelope criterion.

44. A computer readable medium having stored therein instructions for causing a central processing unit to execute the method of Claim 43.

45. The method of Claim 43 wherein the multi-component data signal includes polynucleotide data.

46. The method of Claim 45 wherein polynucleotide data includes any of DNA, cDNA or mRNA data.

47. The method of Claim 43 further comprising visually displaying the third final subset of overlapping data on a display device.

48. A multi-component data signal processing system, comprising in combination:

a data sample and reference calibration module for processing known and target biotechnology samples;

an optional broadband signal collection module for collecting experimental data from laser-induced fluorescence of biotechnology samples;

a filtering and baseline module for removing spectral overlap and normalizing of experimental data;

a reference and calibration module for standard size detection with false peak rejection and clutter removal, data size calibration, envelope detection and data magnitude reducing and data stutter removal of experimental data;

a storage module for storing processed experimental data; and

a display module for visually displaying processed experimental data.

FIG. 1

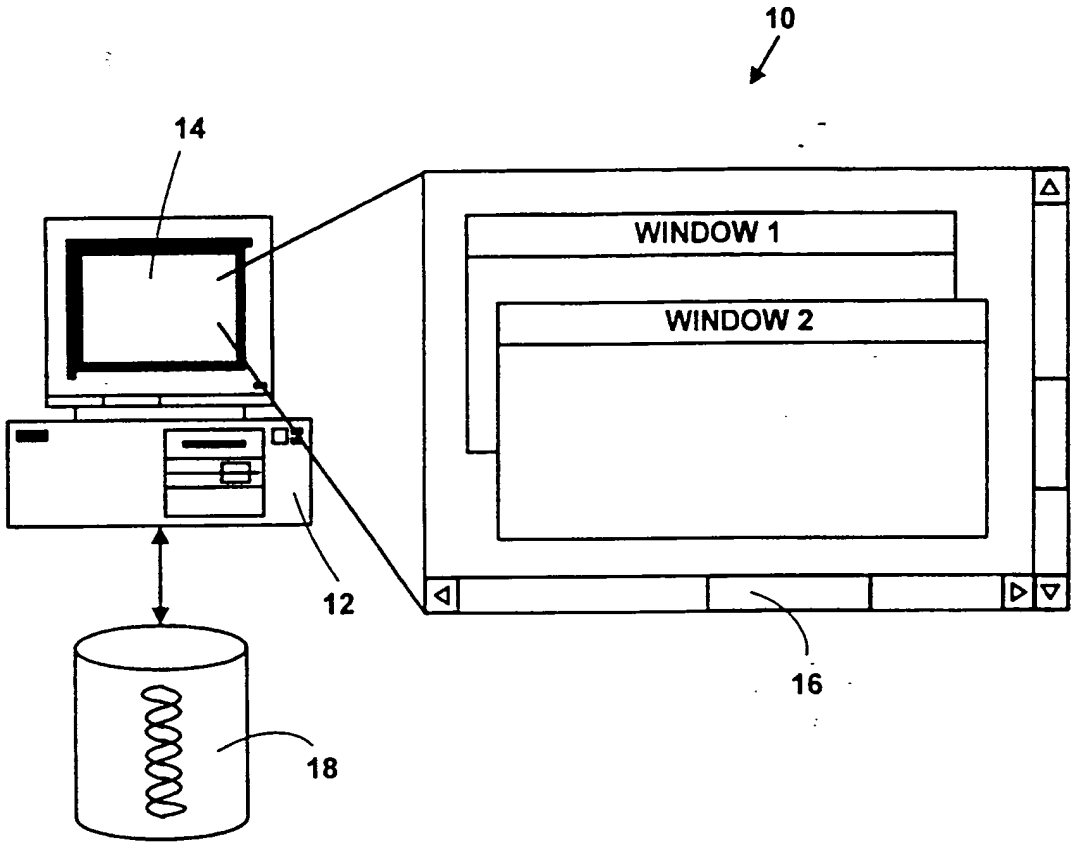


FIG. 2

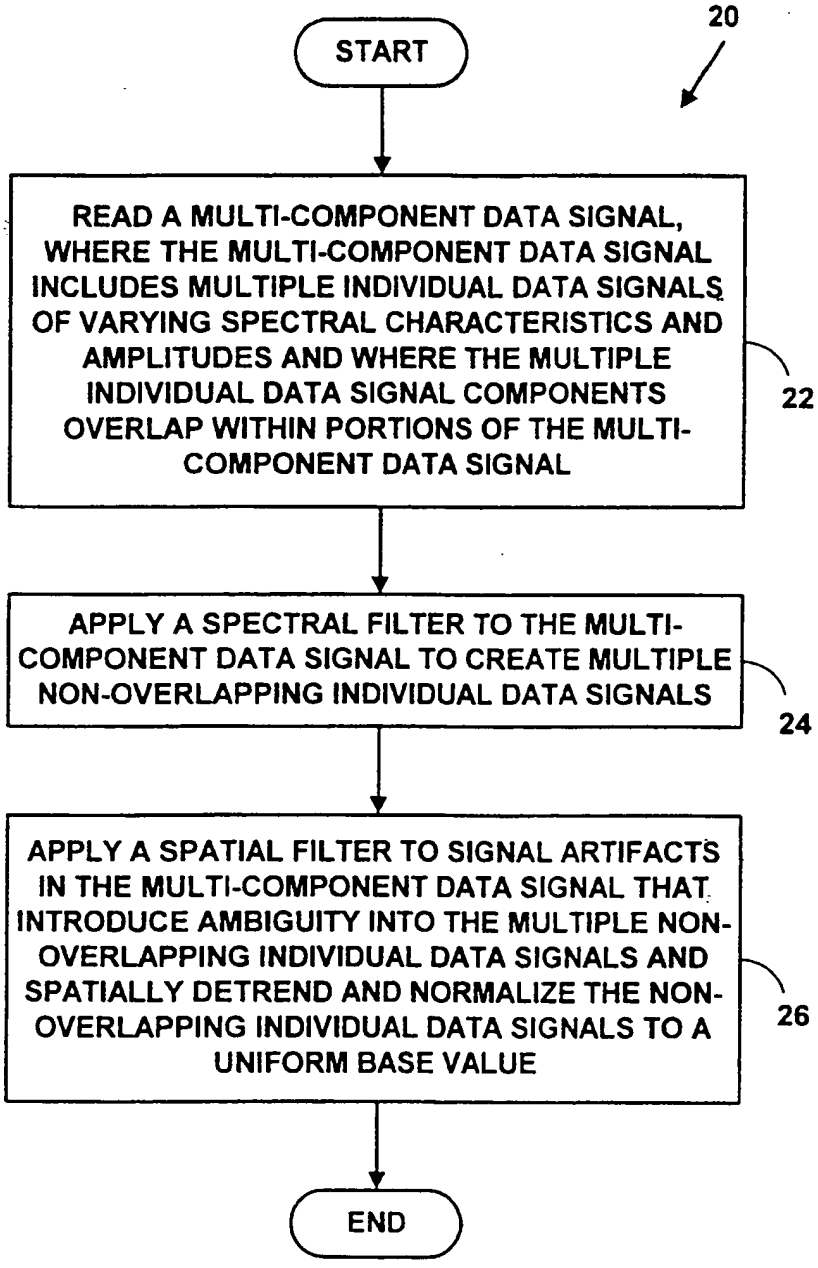

20
START
READ A MULTI-COMPONENT DATA SIGNAL, WHERE THE MULTI-COMPONENT DATA SIGNAL INCLUDES MULTIPLE INDIVIDUAL DATA SIGNALS OF VARYING SPECTRAL CHARACTERISTICS AND AMPLITUDES AND WHERE THE MULTIPLE INDIVIDUAL DATA SIGNAL COMPONENTS OVERLAP WITHIN PORTIONS OF THE MULTI-COMPONENT DATA SIGNAL
22
APPLY A SPECTRAL FILTER TO THE MULTI-COMPONENT DATA SIGNAL TO CREATE MULTIPLE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS
24
APPLY A SPATIAL FILTER TO SIGNAL ARTIFACTS IN THE MULTI-COMPONENT DATA SIGNAL THAT INTRODUCE AMBIGUITY INTO THE MULTIPLE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS AND SPATIALLY DETREND AND NORMALIZE THE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS TO A UNIFORM BASE VALUE
26
END

FIG. 3A

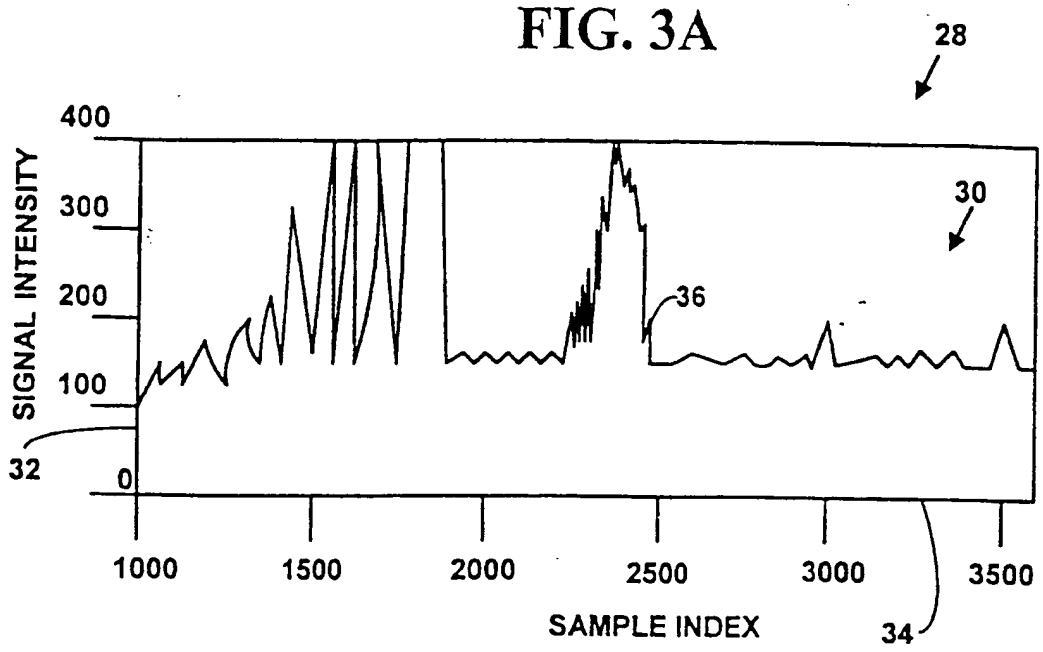

28
400
300
200
100
0
SIGNAL INTENSITY
32
30
36
1000
1500
2000
2500
3000
3500
SAMPLE INDEX
34

FIG. 3D

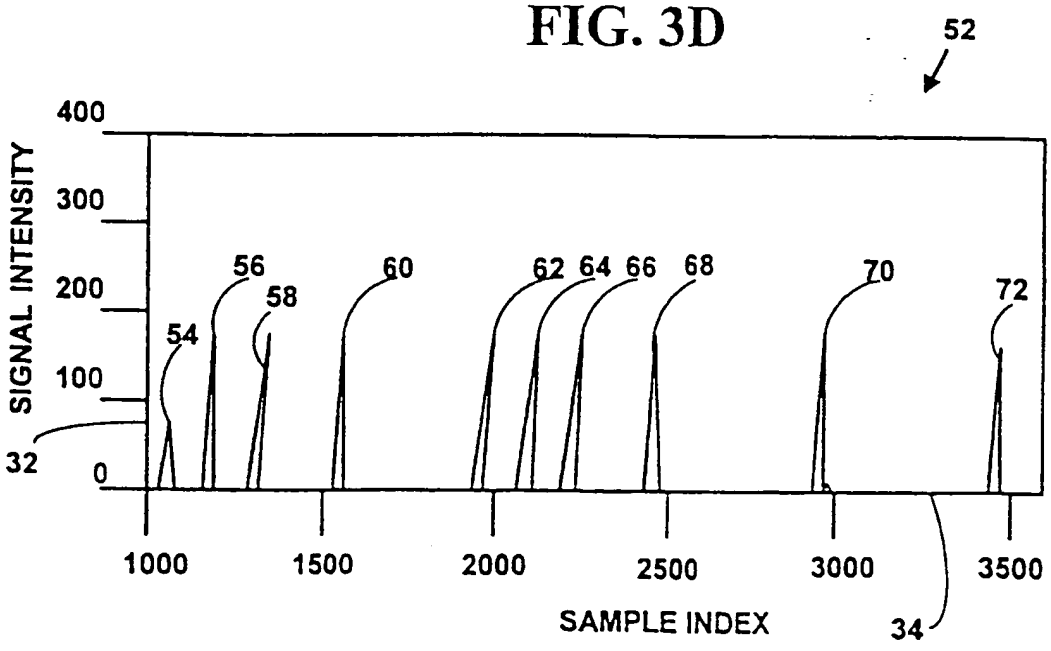

52
400
300
200
100
0
SIGNAL INTENSITY
32
54
56
58
60
62
64
66
68
70
72
1000
1500
2000
2500
3000
3500
SAMPLE INDEX
34

FIG. 3B

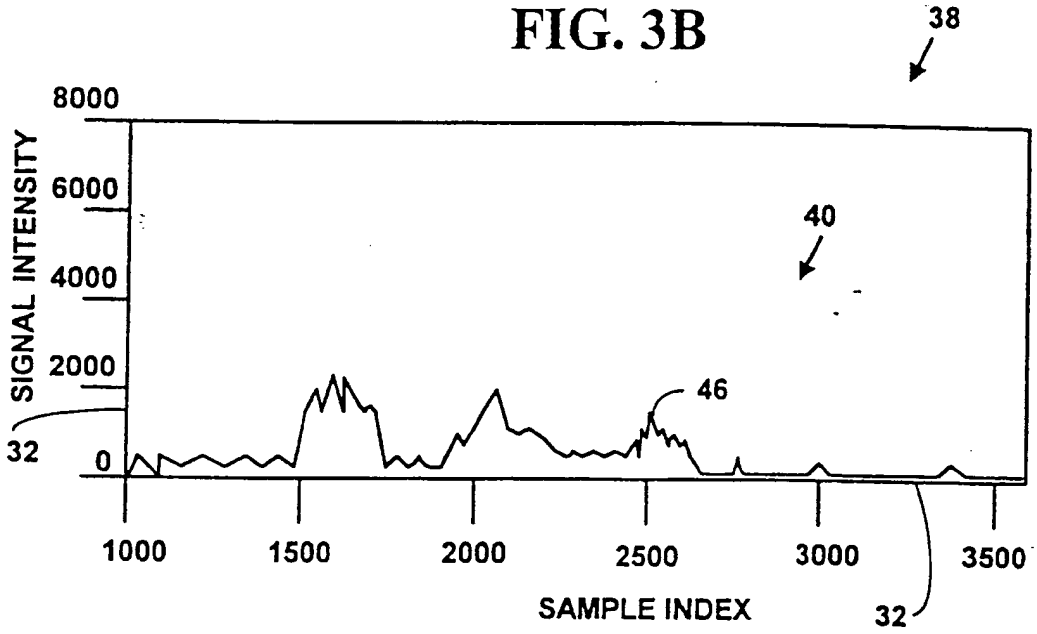

38
8000
6000
4000
2000
0
SIGNAL INTENSITY
32
40
46
1000
1500
2000
2500
3000
3500
SAMPLE INDEX
32

FIG. 3C

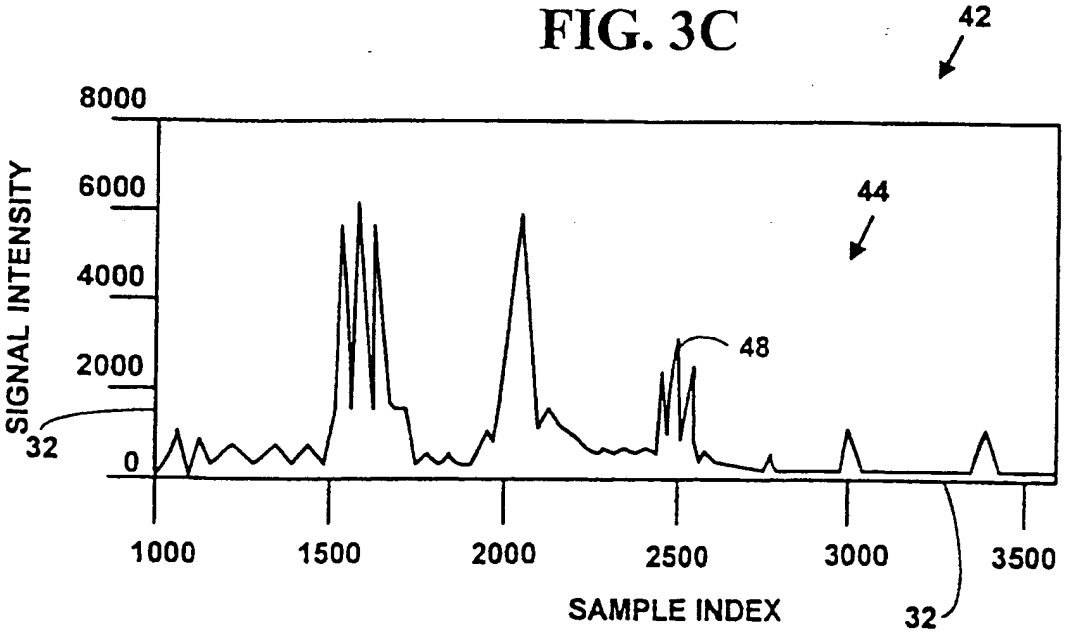

42
8000
6000
4000
2000
0
SIGNAL INTENSITY
32
44
48
1000
1500
2000
2500
3000
3500
SAMPLE INDEX
32

FIG. 4

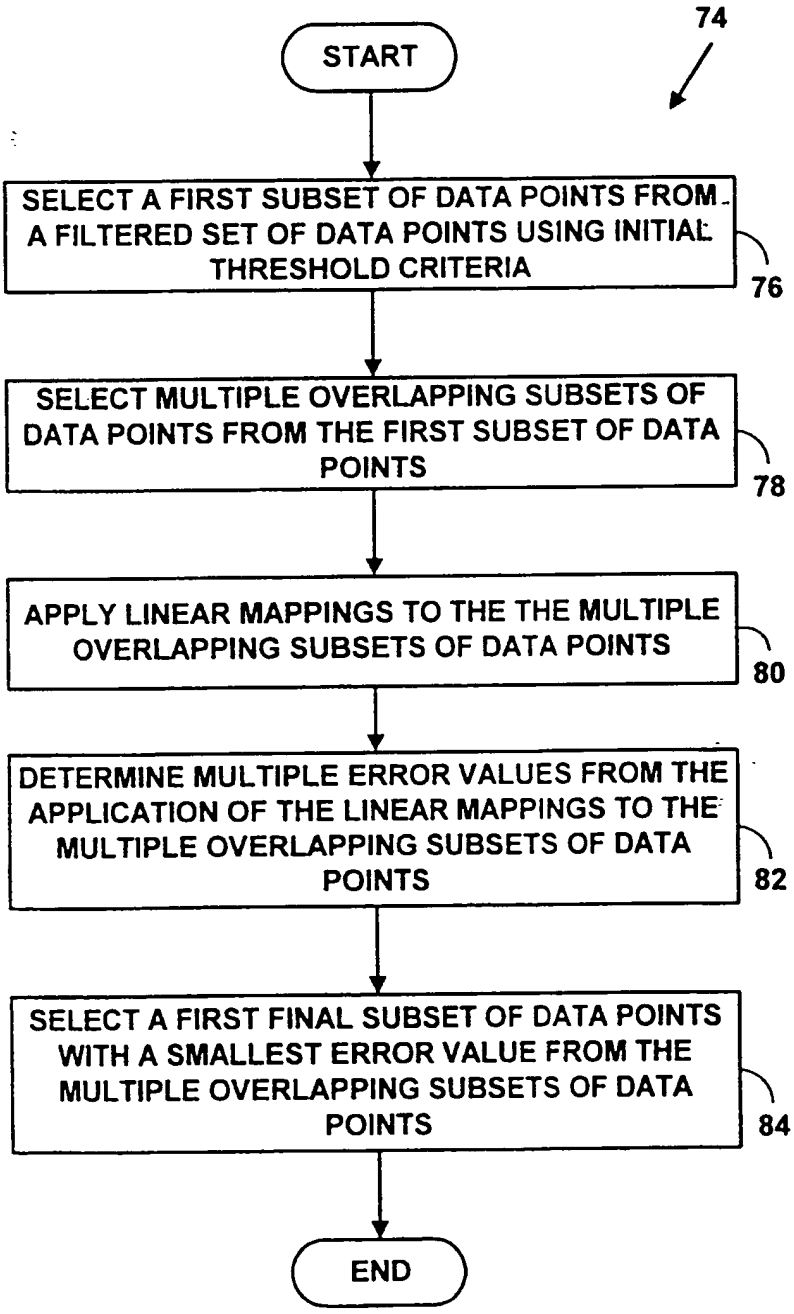

74
START
SELECT A FIRST SUBSET OF DATA POINTS FROM A FILTERED SET OF DATA POINTS USING INITIAL THRESHOLD CRITERIA
76
SELECT MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS FROM THE FIRST SUBSET OF DATA POINTS
78
APPLY LINEAR MAPPINGS TO THE THE MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS
80
DETERMINE MULTIPLE ERROR VALUES FROM THE APPLICATION OF THE LINEAR MAPPINGS TO THE MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS
82
SELECT A FIRST FINAL SUBSET OF DATA POINTS WITH A SMALLEST ERROR VALUE FROM THE MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS
84
END

FIG. 5

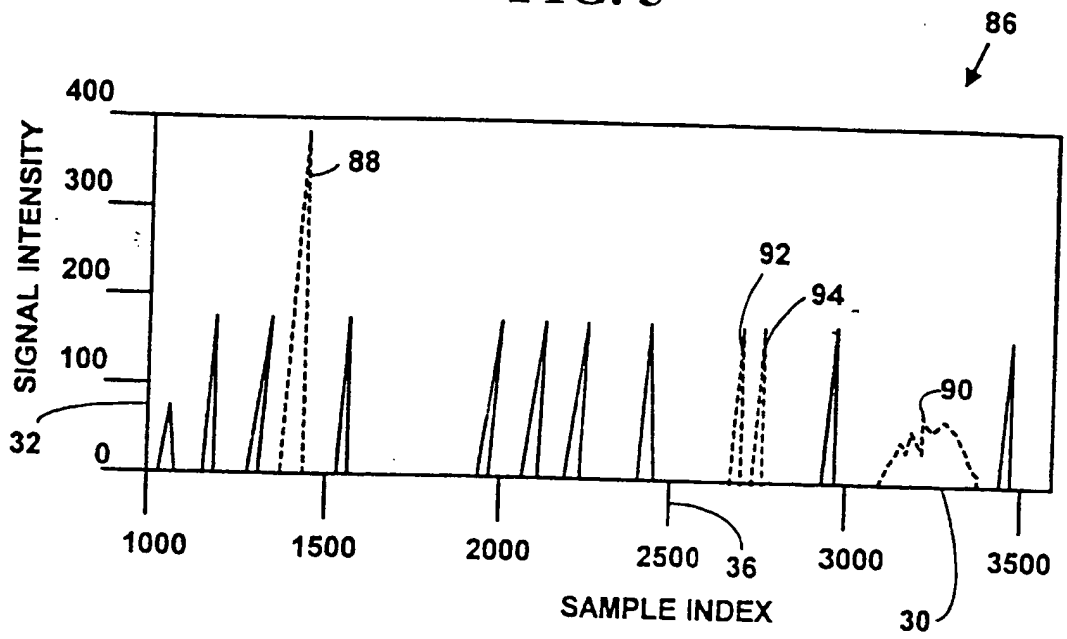

86
400
300
200
100
32
0
SIGNAL INTENSITY
88
92
94
90
1000
1500
2000
2500
36
3000
3500
30
SAMPLE INDEX

FIG. 7

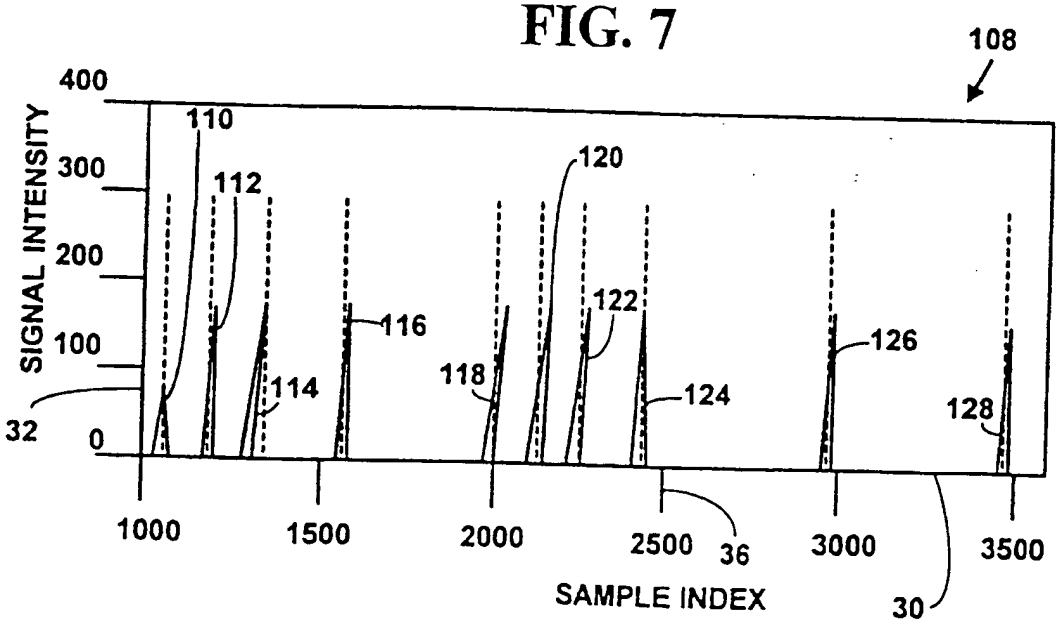

108
400
300
200
100
32
0
SIGNAL INTENSITY
110
112
114
116
118
120
122
124
126
128
1000
1500
2000
2500
36
3000
3500
30
SAMPLE INDEX

FIG. 6

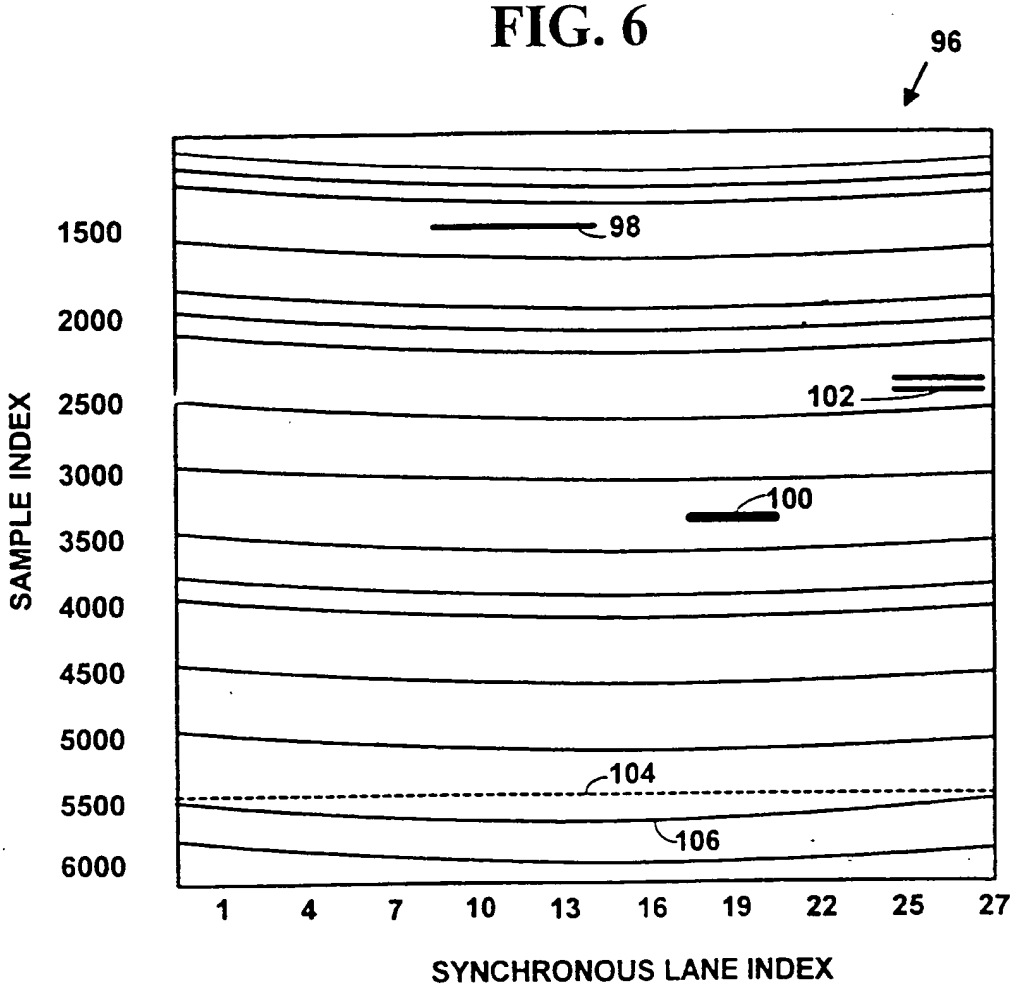

96
98
102
100
104
106
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
SAMPLE INDEX
1
4
7
10
13
16
19
22
25
27
SYNCHRONOUS LANE INDEX

FIG. 8

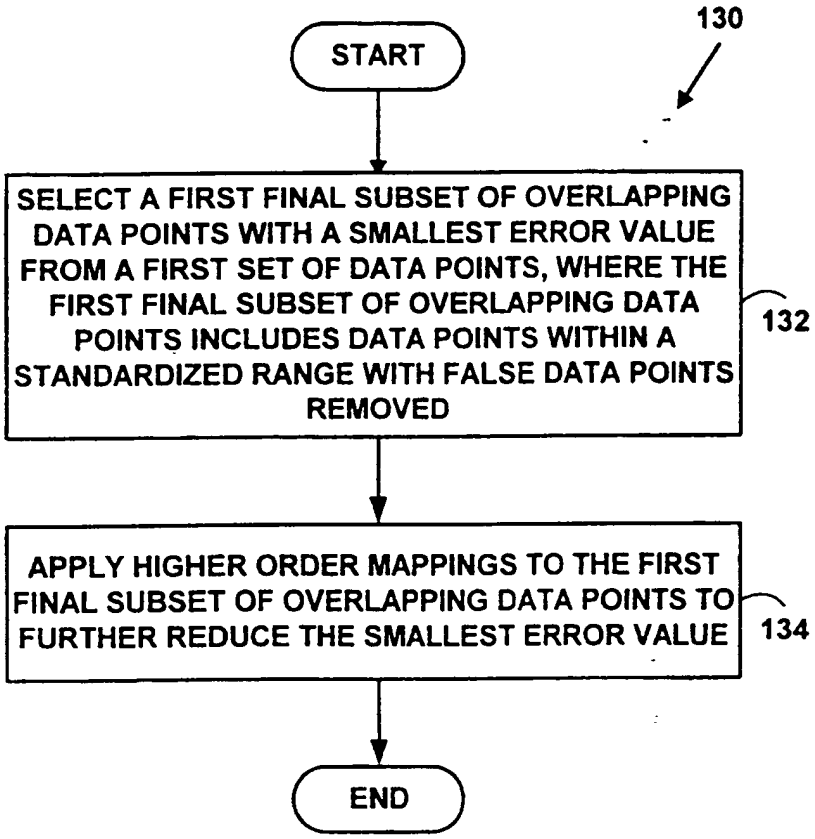

130
START
SELECT A FIRST FINAL SUBSET OF OVERLAPPING DATA POINTS WITH A SMALLEST ERROR VALUE FROM A FIRST SET OF DATA POINTS, WHERE THE FIRST FINAL SUBSET OF OVERLAPPING DATA POINTS INCLUDES DATA POINTS WITHIN A STANDARDIZED RANGE WITH FALSE DATA POINTS REMOVED
132
APPLY HIGHER ORDER MAPPINGS TO THE FIRST FINAL SUBSET OF OVERLAPPING DATA POINTS TO FURTHER REDUCE THE SMALLEST ERROR VALUE
134
END

FIG. 9A

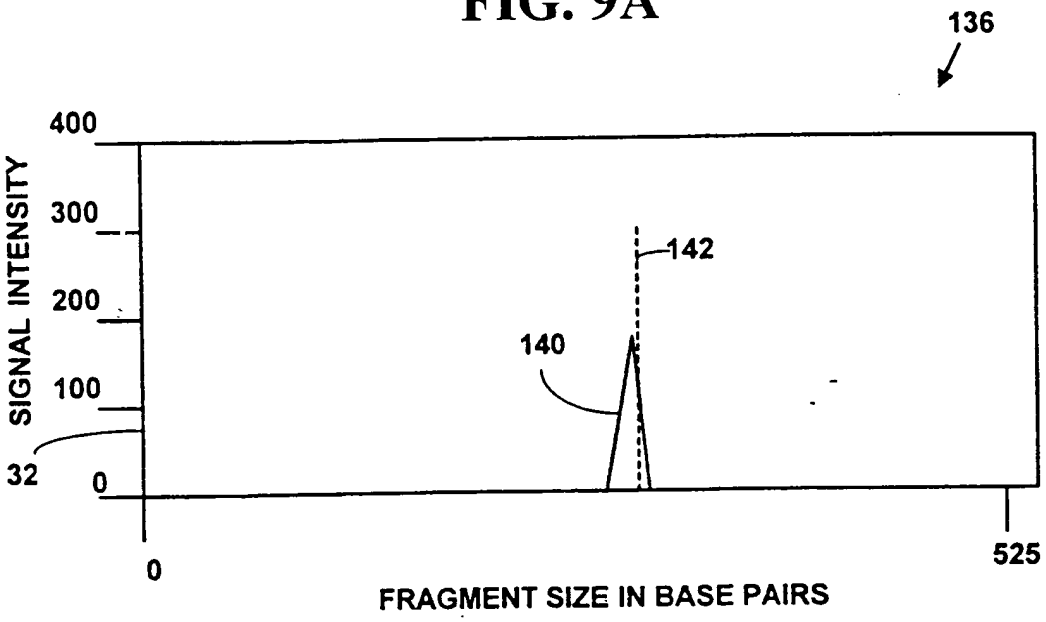

136
400
300
200
100
32
0
SIGNAL INTENSITY
142
140
0
525
FRAGMENT SIZE IN BASE PAIRS

FIG. 9B

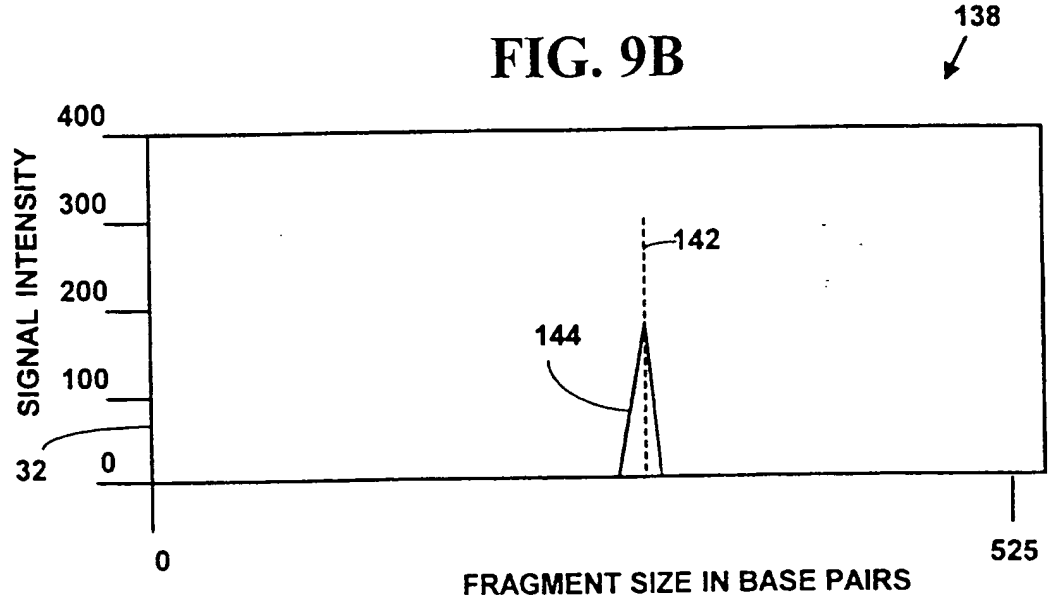

138
400
300
200
100
32
0
SIGNAL INTENSITY
142
144
0
525
FRAGMENT SIZE IN BASE PAIRS

FIG. 10

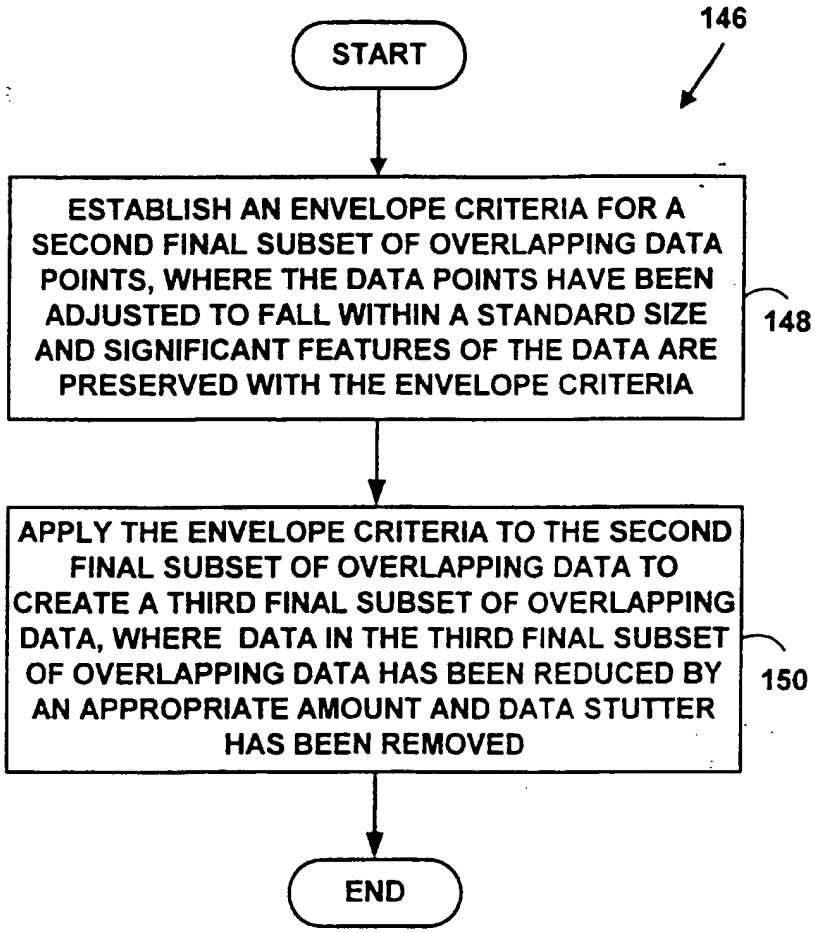

146
START
ESTABLISH AN ENVELOPE CRITERIA FOR A SECOND FINAL SUBSET OF OVERLAPPING DATA POINTS, WHERE THE DATA POINTS HAVE BEEN ADJUSTED TO FALL WITHIN A STANDARD SIZE AND SIGNIFICANT FEATURES OF THE DATA ARE PRESERVED WITH THE ENVELOPE CRITERIA
148
APPLY THE ENVELOPE CRITERIA TO THE SECOND FINAL SUBSET OF OVERLAPPING DATA TO CREATE A THIRD FINAL SUBSET OF OVERLAPPING DATA, WHERE DATA IN THE THIRD FINAL SUBSET OF OVERLAPPING DATA HAS BEEN REDUCED BY AN APPROPRIATE AMOUNT AND DATA STUTTER HAS BEEN REMOVED
150
END

FIG. 11A

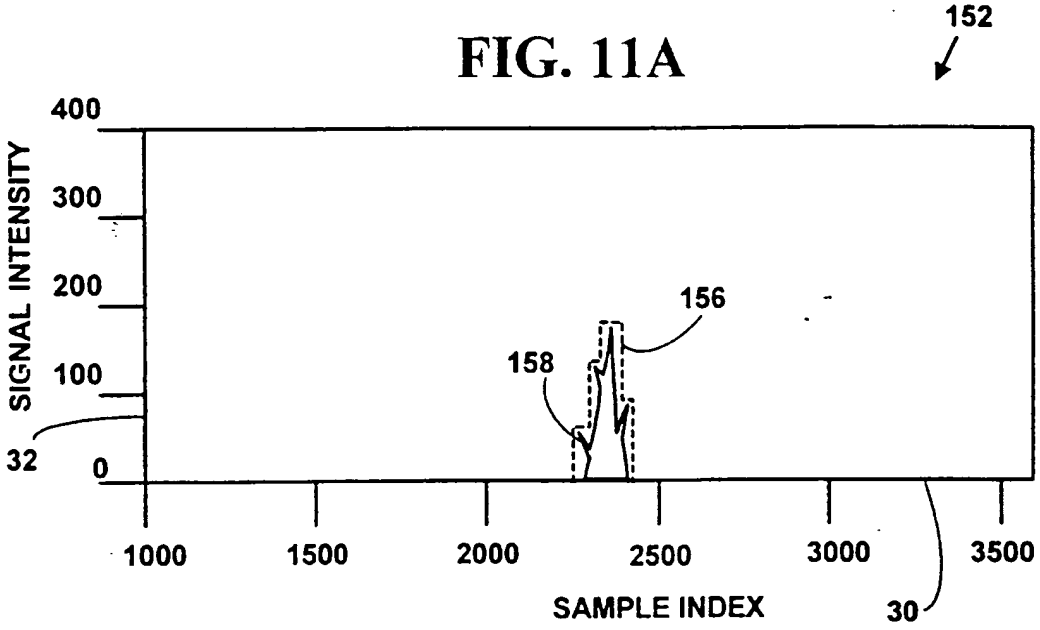


FIG. 11B

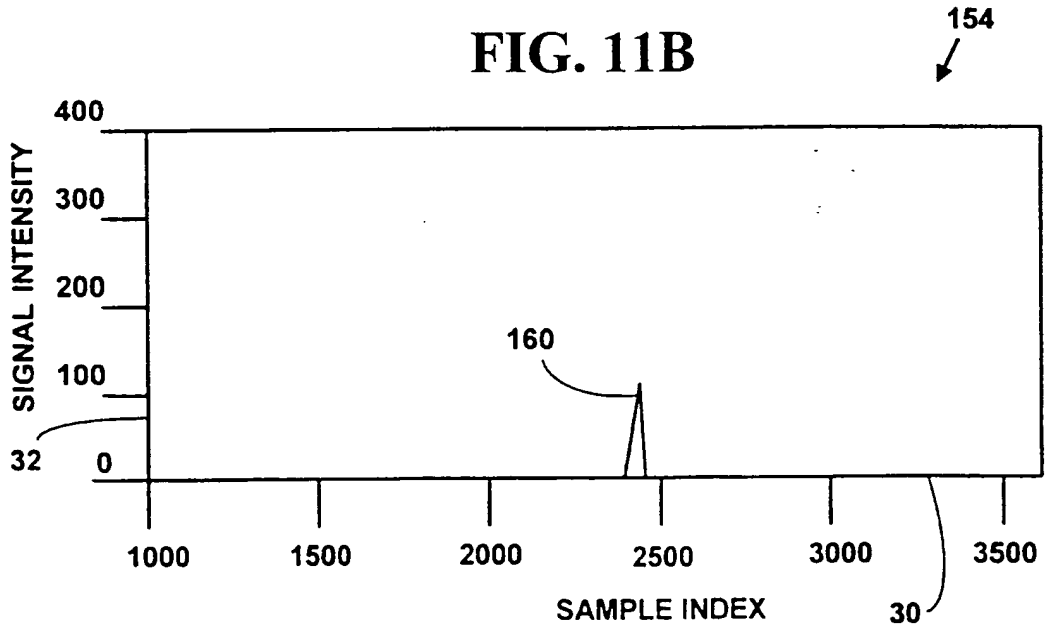


FIG. 12A

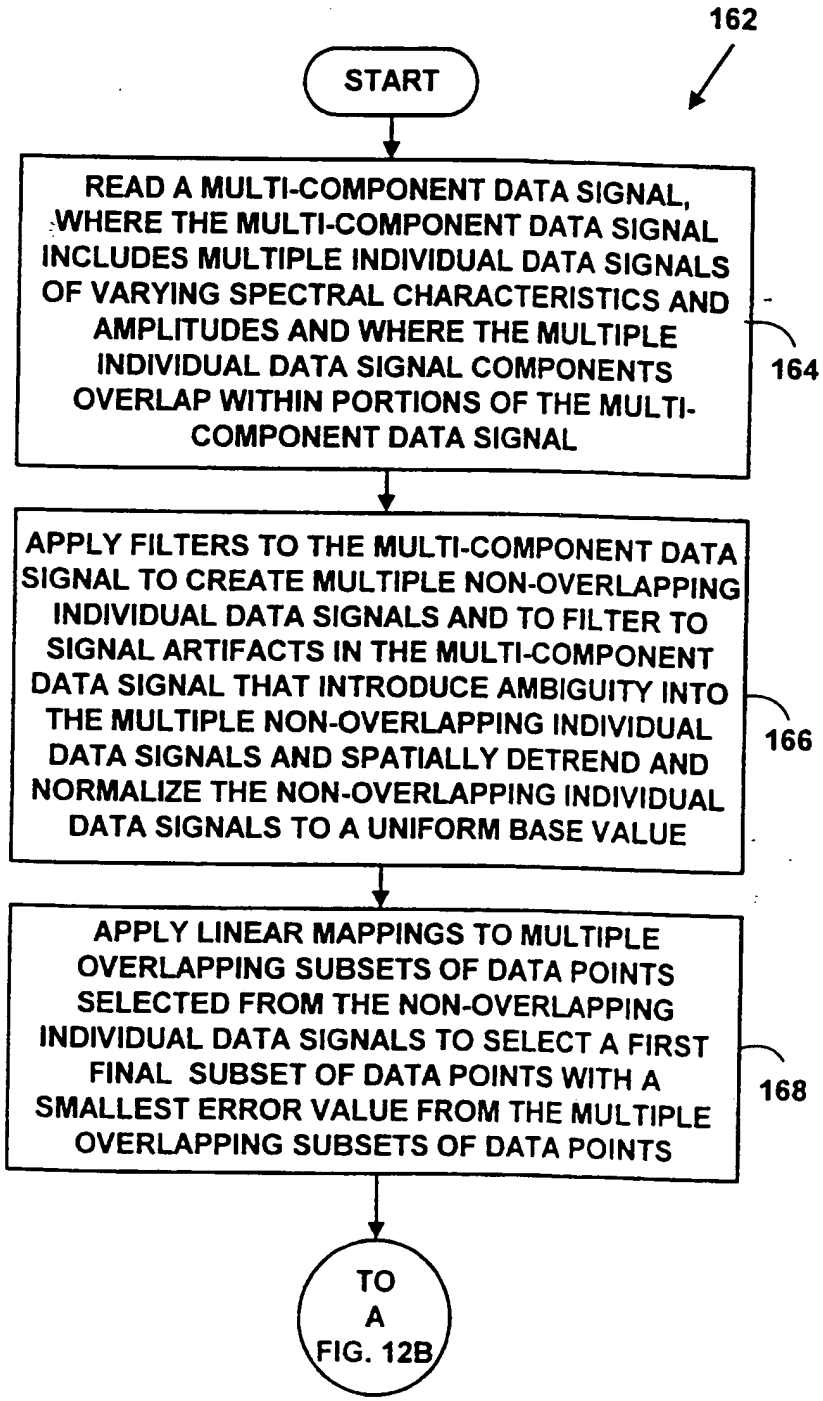

162
START
READ A MULTI-COMPONENT DATA SIGNAL, WHERE THE MULTI-COMPONENT DATA SIGNAL INCLUDES MULTIPLE INDIVIDUAL DATA SIGNALS OF VARYING SPECTRAL CHARACTERISTICS AND AMPLITUDES AND WHERE THE MULTIPLE INDIVIDUAL DATA SIGNAL COMPONENTS OVERLAP WITHIN PORTIONS OF THE MULTI-COMPONENT DATA SIGNAL
164
APPLY FILTERS TO THE MULTI-COMPONENT DATA SIGNAL TO CREATE MULTIPLE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS AND TO FILTER TO SIGNAL ARTIFACTS IN THE MULTI-COMPONENT DATA SIGNAL THAT INTRODUCE AMBIGUITY INTO THE MULTIPLE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS AND SPATIALLY DETREND AND NORMALIZE THE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS TO A UNIFORM BASE VALUE
166
APPLY LINEAR MAPPINGS TO MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS SELECTED FROM THE NON-OVERLAPPING INDIVIDUAL DATA SIGNALS TO SELECT A FIRST FINAL SUBSET OF DATA POINTS WITH A SMALLEST ERROR VALUE FROM THE MULTIPLE OVERLAPPING SUBSETS OF DATA POINTS
168
TO A FIG. 12B

FIG. 12B

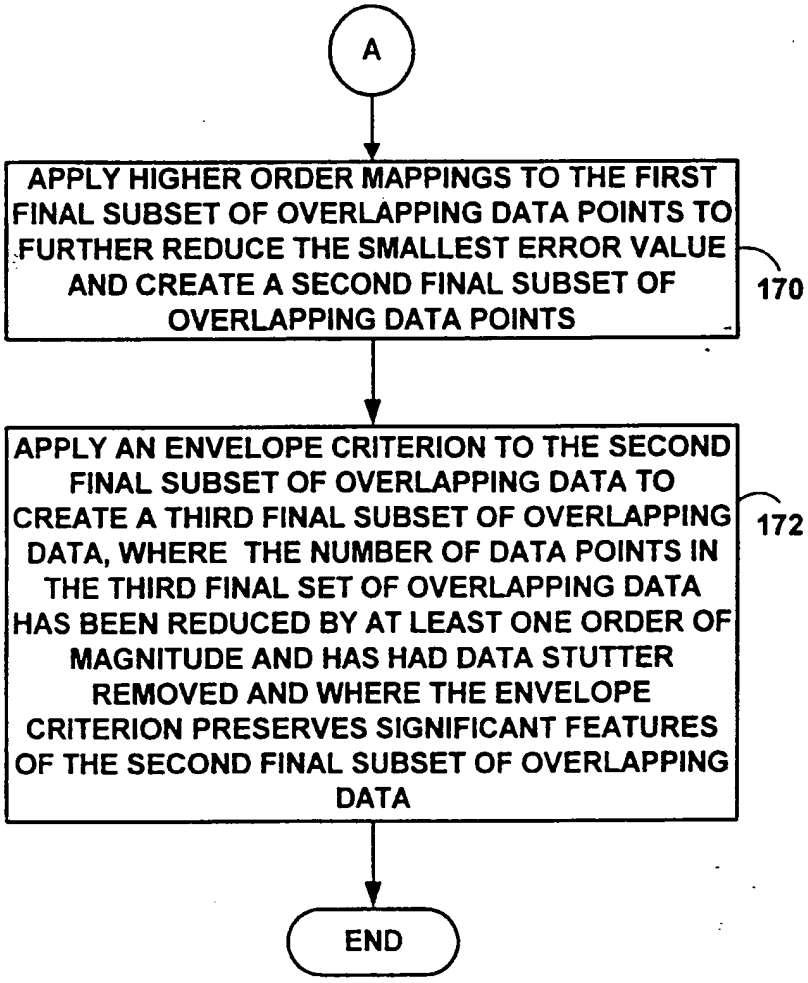

A
APPLY HIGHER ORDER MAPPINGS TO THE FIRST FINAL SUBSET OF OVERLAPPING DATA POINTS TO FURTHER REDUCE THE SMALLEST ERROR VALUE AND CREATE A SECOND FINAL SUBSET OF OVERLAPPING DATA POINTS
170
APPLY AN ENVELOPE CRITERION TO THE SECOND FINAL SUBSET OF OVERLAPPING DATA TO CREATE A THIRD FINAL SUBSET OF OVERLAPPING DATA, WHERE THE NUMBER OF DATA POINTS IN THE THIRD FINAL SET OF OVERLAPPING DATA HAS BEEN REDUCED BY AT LEAST ONE ORDER OF MAGNITUDE AND HAS HAD DATA STUTTER REMOVED AND WHERE THE ENVELOPE CRITERION PRESERVES SIGNIFICANT FEATURES OF THE SECOND FINAL SUBSET OF OVERLAPPING DATA
172
END

FIG. 13A

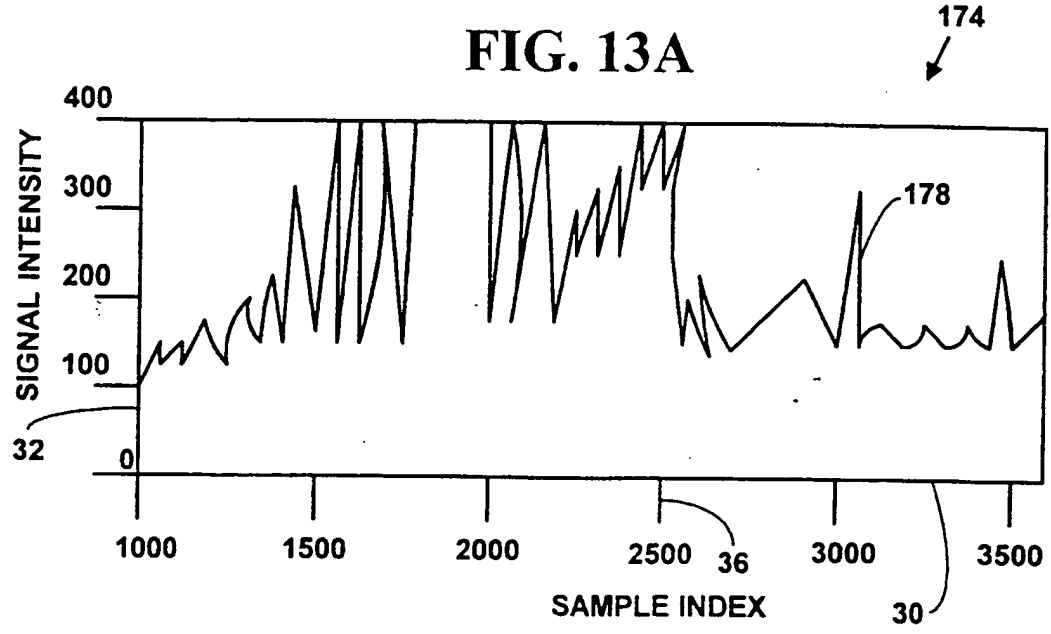


FIG. 13B

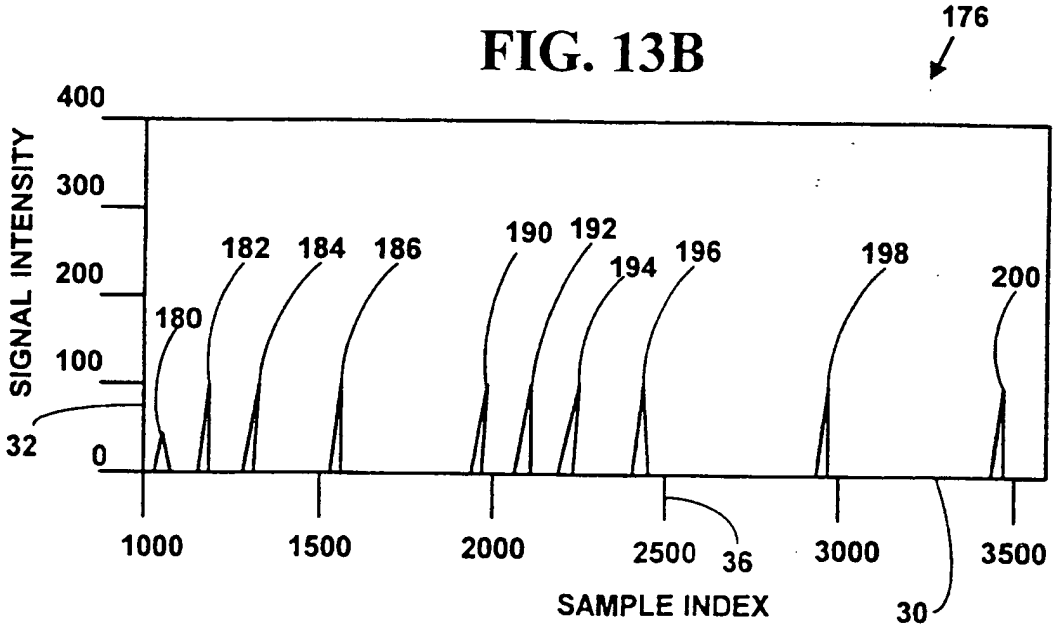
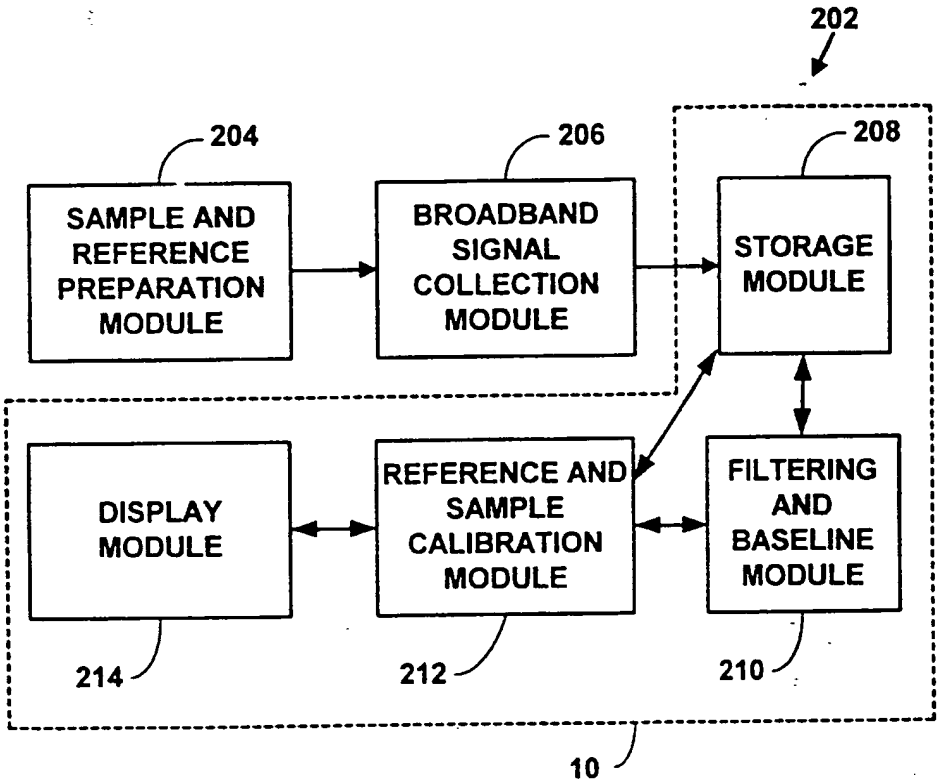


FIG. 14


202
204
206
208
SAMPLE AND REFERENCE PREPARATION MODULE
BROADBAND SIGNAL COLLECTION MODULE
STORAGE MODULE
DISPLAY MODULE
REFERENCE AND SAMPLE CALIBRATION MODULE
FILTERING AND BASELINE MODULE
214
212
210
10

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 November 2000 (30.11.2000)

PCT

(10) International Publication Number
WO 00/72182 A3

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/US00/14159

(22) International Filing Date: 23 May 2000 (23.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/318,699 25 May 1999 (25.05.1999) US

(71) Applicant *(for all designated States except US)*: DIGITAL GENE TECHNOLOGIES, INC. [US/US]; 11149 North Torrey Pines Road, Suite 110, La Jolla, CA 90237 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: GRACE, Dennis, R. [US/US]; 3137 Fenelon Street, San Diego, CA 92106 (US). DURHAM, Jayson, T. [US/US]; 10359 Mountain View Lane, Lakeside, CA 92040 (US).

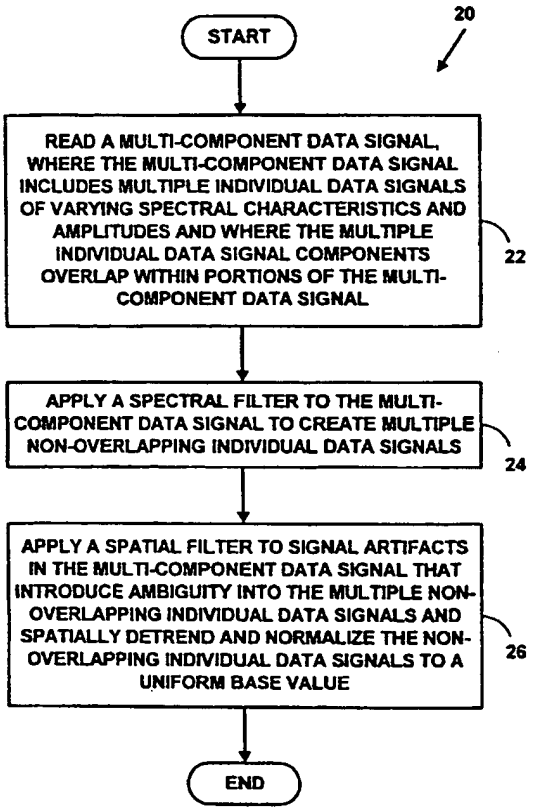
(74) Agent: LESAVICH, Stephen; McDonnell Boehnen Hulbert & Berghoff, Suite 3200, 300 South Wacker Drive, Chicago, IL 60606 (US).

(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHODS AND SYSTEM FOR AMPLITUDE NORMALIZATION AND SELECTION OF DATA PEAKS



(57) Abstract: Methods and system for amplitude normalization and selection of data peaks from experimental data including polynucleotide data such as DNA, cDNA or mRNA from biotechnology experiments. The methods and systems include removing spectral overlap, spatially detrending and normalizing a multi-component data signal into experimental data from a desired experiment (e.g., biotechnology data). Standard data sizes are determined and data clutter is rejected for filtered experimental data. Data sizes are calibrated and error removed from experimental data. Data stutter is removed and the number of data values is reduced. The methods and system help automate the processing of experimental data to eliminate or reduce errors and leave processed experimental data in a format suitable for visual display, comparative analysis and other analysis. The methods and systems may help reduce or eliminate inconsistencies in processing experimental data that typically lead to unreliable or erroneous results. The methods and system of the present invention may be used to refine processing of biotechnology data with new techniques that can be used for bioinformatics and for other types of experimental data that are visual displayed (e.g., telecommunications data, electrical data for electrical devices, optical data, physical data, or other data).

WO 00/72182 A3

Published:
— *With international search report.*

(88) Date of publication of the international search report:
15 March 2001

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

Inter. ...nal Application No
PCT/US 00/14159

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, IBM-TDB, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 98 00708 A (CHI VRIJMOED ;GILCHRIST RODNEY D (CA); VISIBLE GENETICS INC (CA)) 8 January 1998 (1998-01-08) abstract; claims 1-3 page 6, line 14 -page 7, line 24	1-48
P,X	WO 00 22173 A (ALON URI ;LEVINE ARNOLD J (US); UNIV PRINCETON (US)) 20 April 2000 (2000-04-20) abstract; claims 41-47; figures 3,4 page 6, line 27 -page 7, line 32	1-48

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
7 December 2000	14/12/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Filloy García, E

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter. .nal Application No
PCT/US 00/14159

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	GIDDINGS M C ET AL: "AN ADAPTIVE, OBJECT ORIENTED STRATEGY FOR BASE CALLING IN DNA SEQUENCE ANALYSIS" NUCLEIC ACIDS RESEARCH,OXFORD UNIVERSITY PRESS, SURREY,GB, vol. 21, no. 19, 1993, pages 4530-4540, XP000919267 ISSN: 0305-1048 the whole document	1-48
A	US 5 853 979 A (DEE GREGORY ET AL) 29 December 1998 (1998-12-29) abstract; claims 1,2; figures 1,2 column 8, line 10 - line 33 column 10, line 39 - line 54	1-48
A	VERBEEK P W ET AL: "2-D ADAPTIVE SMOOTHING BY 3-D DISTANCE TRANSFORMATION" PATTERN RECOGNITION LETTERS,NL,NORTH-HOLLAND PUBL. AMSTERDAM, vol. 9, no. 1, 1989, pages 53-65, XP000098457 ISSN: 0167-8655	
A	LUCKE L ET AL: "A DIGIT-SERIAL ARCHITECTURE FOR GRAY-SCALE MORPHOLOGICAL FILTERING" IEEE TRANSACTIONS ON IMAGE PROCESSING,US,IEEE INC. NEW YORK, vol. 4, no. 3, 1 March 1995 (1995-03-01), pages 387-391, XP000501913 ISSN: 1057-7149	

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inten nal Application No
PCT/US 00/14159

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9800708 A	08-01-1998	US 5916747 A	29-06-1999
		AU 3250797 A	21-01-1998
		CA 2259314 A	08-01-1998
		US 5981186 A	09-11-1999
WO 0022173 A	20-04-2000	AU 1597400 A	01-05-2000
US 5853979 A	29-12-1998	AU 700410 B	07-01-1999
		AU 6403996 A	05-02-1997
		CA 2225385 A	23-01-1997
		DE 69601720 D	15-04-1999
		DE 69601720 T	22-07-1999
		EP 0835442 A	15-04-1998
		JP 11509622 T	24-08-1999
		WO 9702488 A	23-01-1997
		US 5834189 A	10-11-1998
		US 5916747 A	29-06-1999
		US 5981186 A	09-11-1999